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 who is in the business of manufacturing and selling high end sales goods wants to send catalogues to their mailing list of 250 new customers and wants to predict the profit that can make by sending the catalogue to these new 250 customers and also wants to take a call based on the expected profit if at all these catalogues needs to be send or not. If the expected profit is < $ 10,000 than the company would take a decision of not sending the catalogues.

2. What data is needed to inform those decisions?

The following are the data needed to inform those decisions:

a) Average Sales (Data provided)
b) Expected profit
c) Customers Average number of Products purchased
d) Profit Margin (50%)- (Data Provided)
e) Catalogue costing (Data Provided)
f) We have the past data about sales from the dataset P1-customers and will build the model. We shall run the score tool between the built model and P1-mailing list to predict the profit for the 250 customers. The Predicted sales when multiplied by the probability that the customer will respond to a catalogue (Score_yes), minus the costing of 250 catalogues will give the profit based on which the company can take a call.

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.

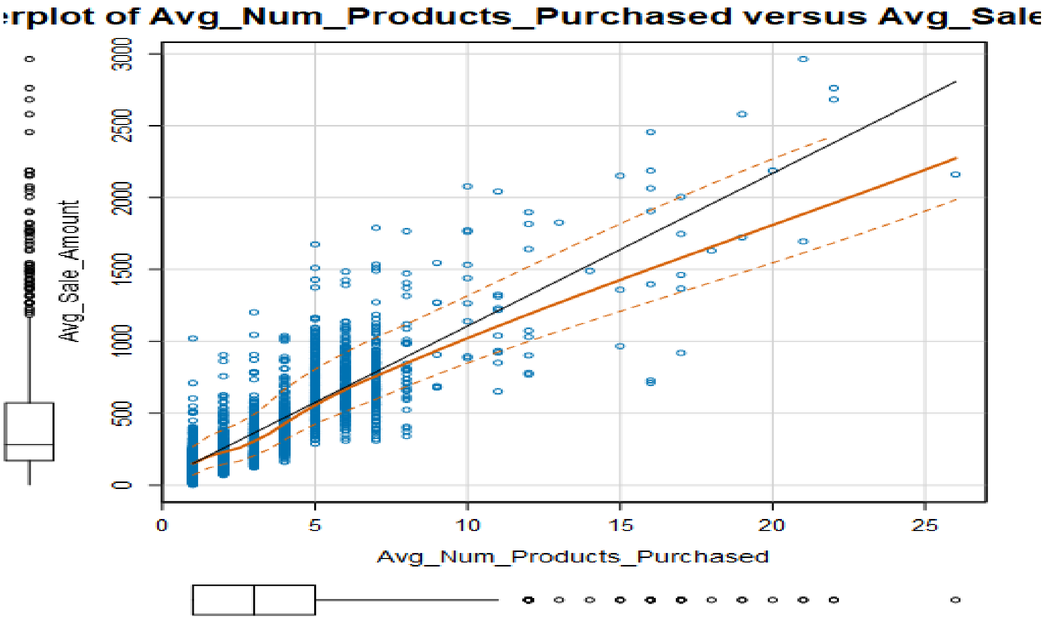
At the minimum, answer these questions:

1. How and why did you select the predictor variables (see supplementary text) in your model? You must explain how your continuous predictor variables you've chosen have a linear relationship with the target variable. Please refer to this lesson to help you explore

your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer.

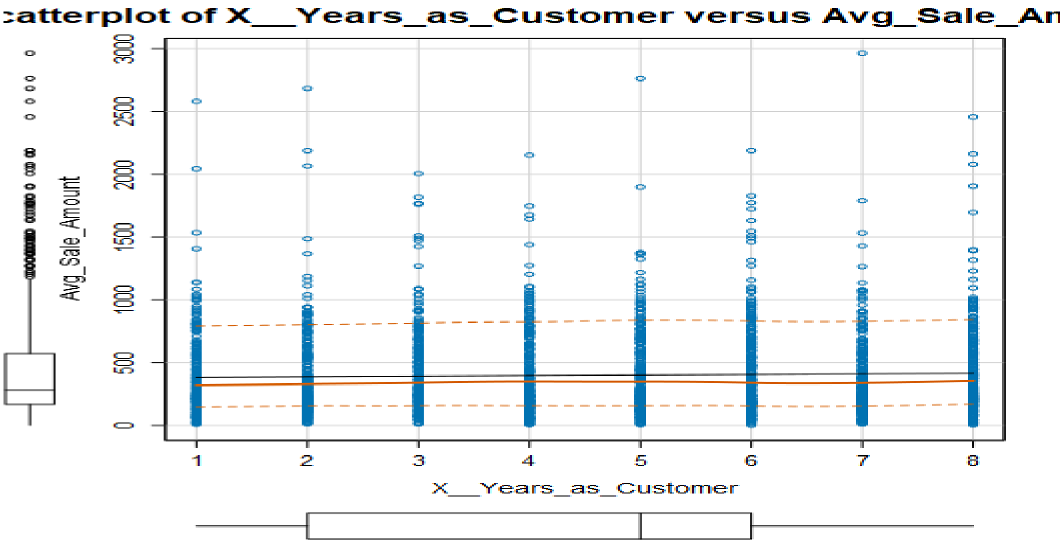
The following are the scatter plots between the continuous predictor variables and the target variable to check if there exists any linear relationship.

1)



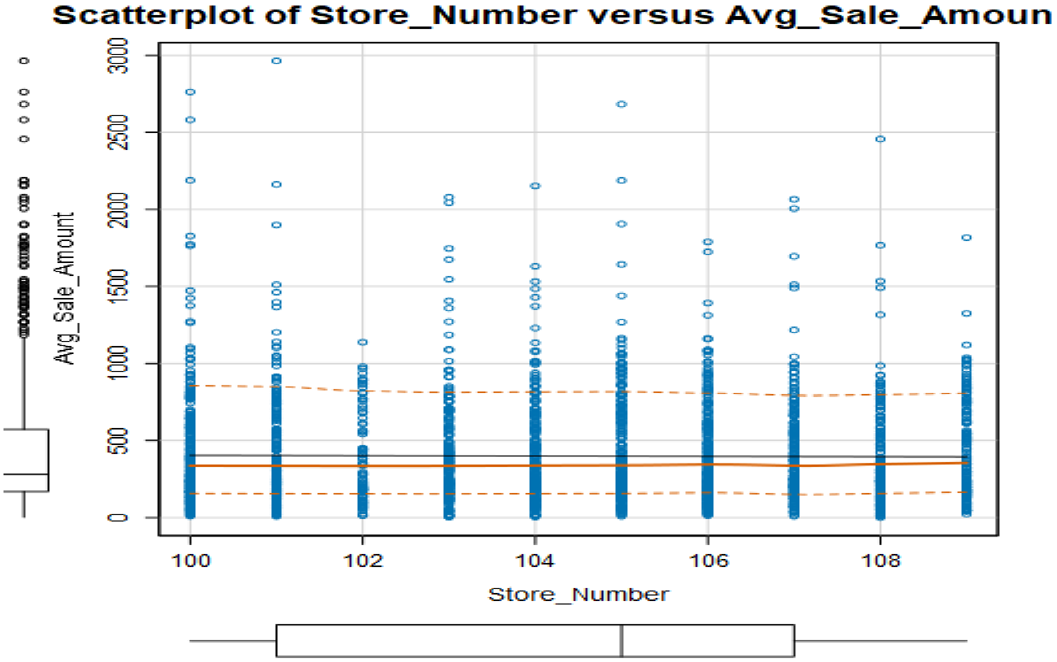
There exists a linear relationship between Average_Num_Products_Purchased and Average_Sales.

2)



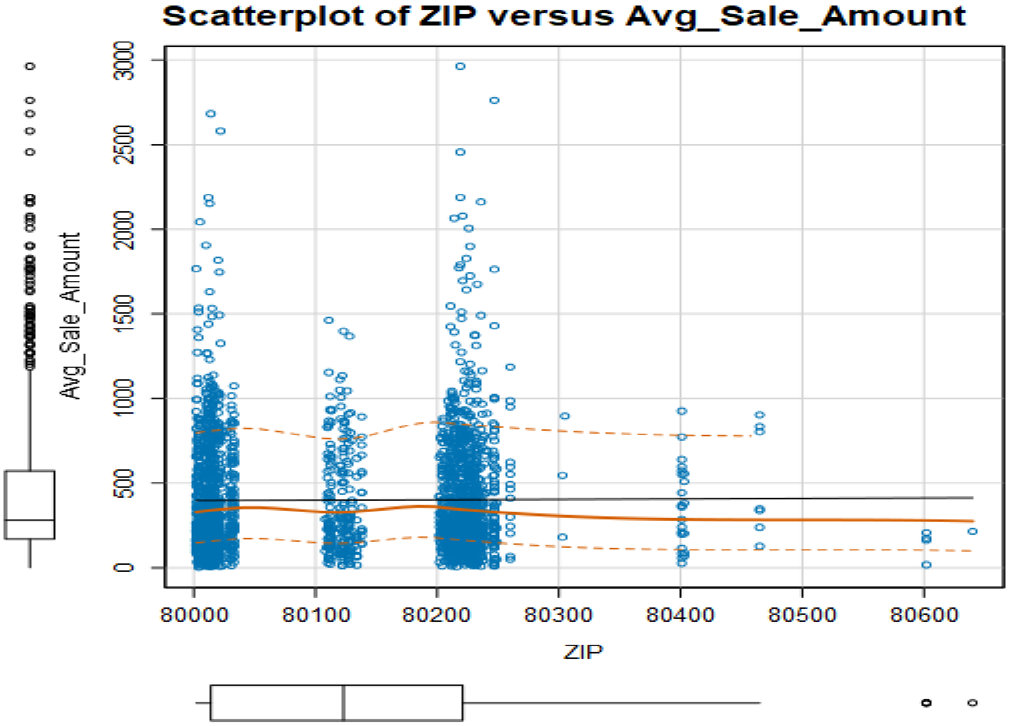
Relationship not linear

3)



Relationship not Linear

4)



Relationship not Linear

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.

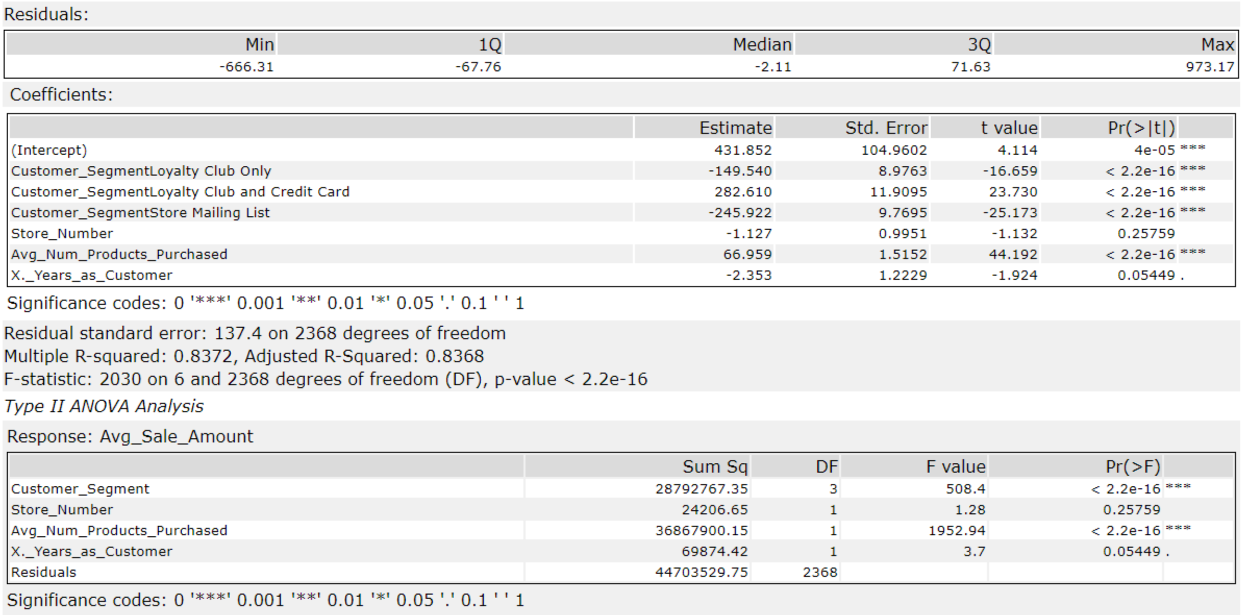
The linear model is a good model. The following are the justifications.

Step -1

From the scatter plots it was found that of all the continuous predictor variables only Avg_number_products_purchased shows a significant linear relationship with the target variable.

Step – 2

A regression model was run with all the other variables to check if there exists any linear relationship with the target variable. Respond_to_last_catalog has been omitted because this variable is not present in the P1-mailing list and does not make any sense (as they are new customers). P-values were used to check if there exists any significant relation with the categorical variables.



Residuals:

Min	1Q	Median	3Q	Max
-666.31	-67.76	-2.11	71.63	973.17

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	431.852	104.9602	4.114	4e-05 ***
Customer_SegmentLoyalty Club Only	-149.540	8.9763	-16.659	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	282.610	11.9095	23.730	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-245.922	9.7695	-25.173	< 2.2e-16 ***
Store_Number	-1.127	0.9951	-1.132	0.25759
Avg_Num_Products_Purchased	66.959	1.5152	44.192	< 2.2e-16 ***
X._Years_as_Customer	-2.353	1.2229	-1.924	0.05449 .

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

Residual standard error: 137.4 on 2368 degrees of freedom
Multiple R-squared: 0.8372, Adjusted R-Squared: 0.8368
F-statistic: 2030 on 6 and 2368 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	28792767.35	3	508.4	< 2.2e-16 ***
Store_Number	24206.65	1	1.28	0.25759
Avg_Num_Products_Purchased	36867900.15	1	1952.94	< 2.2e-16 ***
X._Years_as_Customer	69874.42	1	3.7	0.05449 .
Residuals	44703529.75	2368		

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

The following linear relationship was found between:

Avg_Num_Products_Purchaed and Avg_sales (Significant Relationship) P value <0.05 ***
Customer_Segment and Avg_sales (Significant Relationship) P value < 0.05***
Store_Number (No Relationship)

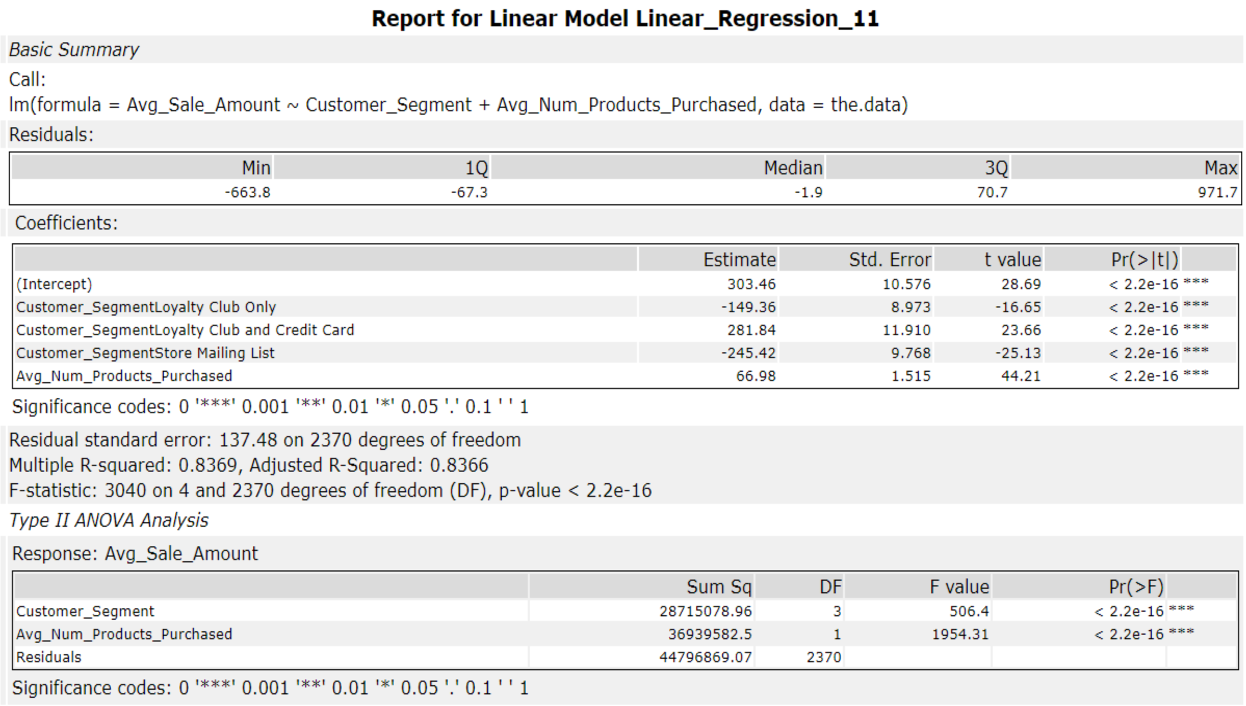
Step 3

After removing the predictor variables which are not significant only 2 predictor variables

a) Customer_segment (Categorical Predictor Variable)

b) Avg_num_Products_purchased (Numeric Predictor Variable)

Were used to build the linear model as shown below.



Report for Linear Model Linear_Regression_11

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

- Both the predictor variables Customer_segment and Avg_num_Products_purchased the P-value (Probability that the coefficient is going to be 0) is very less. A low P-value of <0.5 is significant to the model because changes to the predictor value are related to the changes to the response variable
- This model is very strong as the R-value (0.8369) is very high and also the R-Square (0.8366) is also high which shows the predictor variables are close to the actual values. R square is statistical measure of how close the data are to the fitted regression line. R-Square values lies between 0-1 with 1 suggesting all variability of the response variable around its mean. So closer to 1 the better.
- So Low P values and High R-Squared values suggests that the model is highly predictive.

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= 303.46 – 149.36 X (customer_segmentLoyalty Club only) + 281.84 X (Customer_SegmentLoyalty Club and Credit Card) – 245.42 X (Customer_SegmentStore mailing list) + 0 X (Customer_segmentCredit Card Only) + 66.98 X (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:

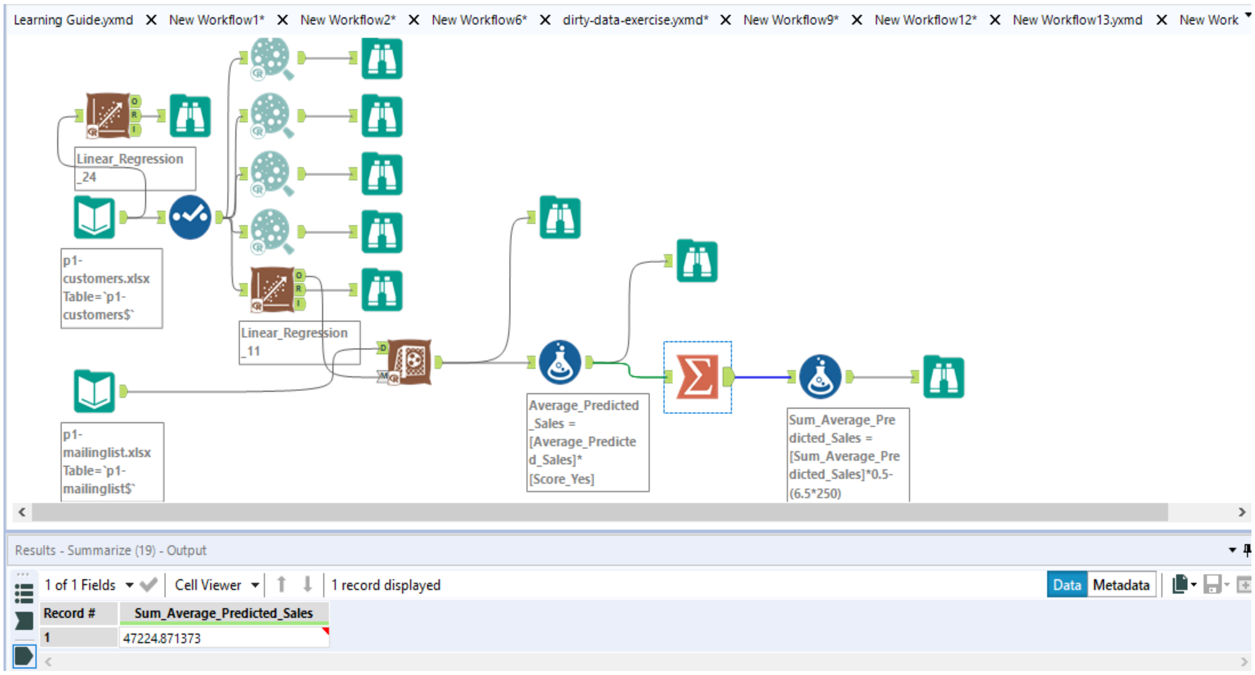
1. What is your recommendation? Should the company send the catalog to these 250 customers?

My recommendation is the company should go ahead with sending the catalogue to these 250 customers as it has met the criteria that being looked into that is the expected profit should be greater than $10000.

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

The following is the process:

a) We ran the scatter plots for the continuous Predictor variables in Alteryx. (Workflow Attached below)
b) From the scatter plots it was found that out of all the continuous predictor variables only Avg_number_products_purchased shows a significant linear relationship with the target variable using the P value. Also checked the R-value to see how good the model is.
c) Some of the other variables like Name, Customer ID, Address, Zip are not significant and hence not used.
d) Calculated the Avg_Predicted_sales by running the score tool to fit the Linear model with the mailing list and using formula tool multiplied the Score_yes and Summarize tool to achieve the Sum_ Average_Predicted_Sales


Learning Guide.yxmd X New Workflow1* X New Workflow2* X New Workflow6* X dirty-data-exercise.yxmd* X New Workflow9* X New Workflow12* X New Workflow13.yxmd X New Work
Linear_Regression_24
p1-customers.xlsx Table=`p1-customers$`
Linear_Regression_11
p1-mailinglist.xlsx Table=`p1-mailinglist$`
Results - Summarize (19) - Output
1 of 1 Fields
Cell Viewer
Data
Metadata

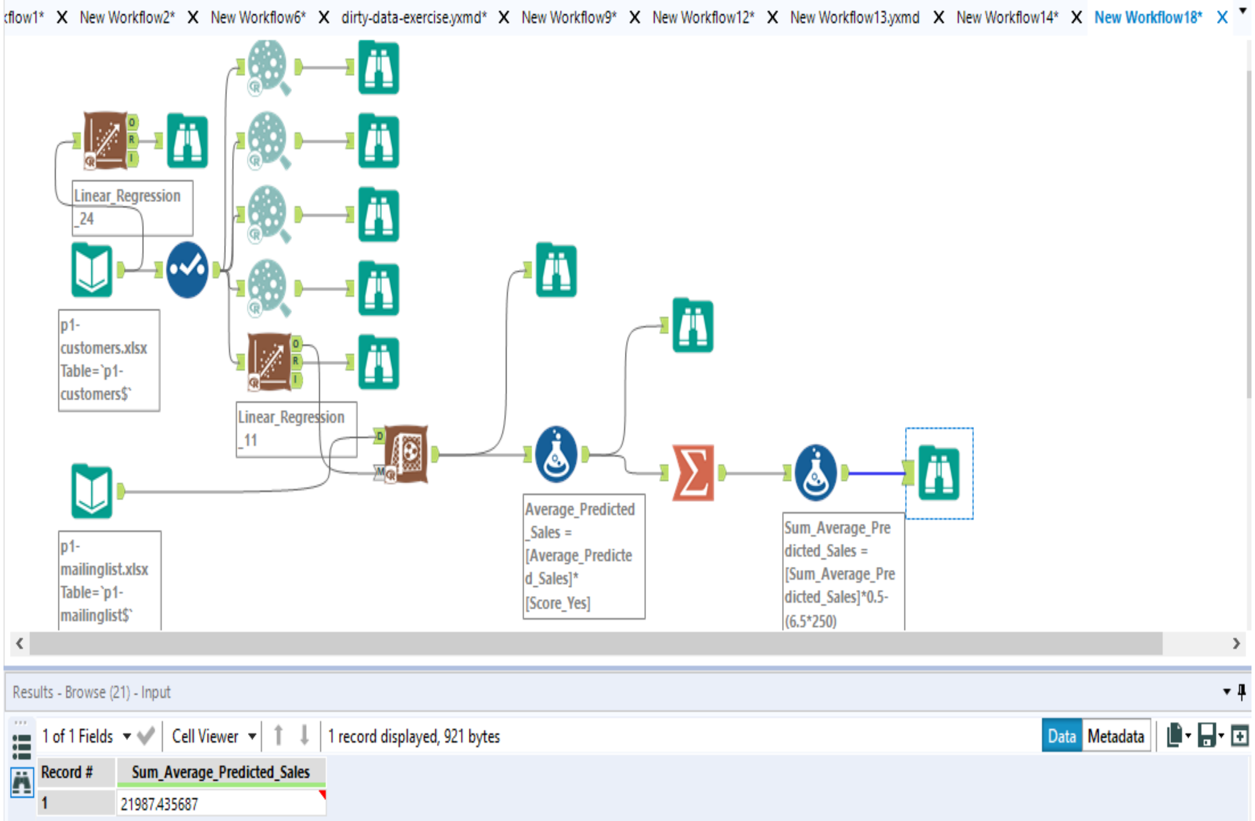
Average_Predicted_Sales = [Average_Predicted_Sales]*[Score_Yes]

Sum_Average_Predicted_Sales = [Sum_Average_Predicted_Sales]*0.5-(6.5*250)

1 record displayed

Record #	Sum_Average_Predicted_Sales
1	47224.871373

e) Calculated the profit by the formula Sum_Average_predicted_sales * 0.5 (profit margin) – (6.5*250)


flow1* X New Workflow2* X New Workflow6* X dirty-data-exercise.yxmd* X New Workflow9* X New Workflow12* X New Workflow13.yxmd X New Workflow14* X New Workflow18* X
Linear_Regression_24
p1-customers.xlsx Table=`p1-customers$`
Linear_Regression_11
p1-mailinglist.xlsx Table=`p1-mailinglist$`
Results - Browse (21) - Input
1 of 1 Fields
Cell Viewer
Data
Metadata

Average_Predicted_Sales = [Average_Predicted_Sales]*[Score_Yes]

Sum_Average_Predicted_Sales = [Sum_Average_Predicted_Sales]*0.5-(6.5*250)

1 record displayed, 921 bytes

Record #	Sum_Average_Predicted_Sales
1	21987.435687

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

Sum_Average_predicted_sales = $47,224.87

Expected Profit = (Sum_Average_predicted_sales X Profit Margin) – (Cost per catalogue X No. of catalogues)
= (47,224.87 X 0.5) – ($6.5 X 250)
= $23,612.44 X $1625 = $21,987.44

The Expected Profit Contribution of $21,987.44 is significantly higher than the limit set of $10,000. Hence the company should go ahead in sending the catalogue to the 250 customers.

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.