MAKE A COPY

Project 1: Predicting Catalog Demand

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/c0b53068-1239-4f01-82bf-24886872f48e/project

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?

Should the company spend the money for sending out the catalog to the new customers (250 new customers)? The management will only send out the catalogs if we estimate a profit of 10,000$. There is an investment for sending out a catalog: Printing and sending costs per catalog: 6.50$ Will the new customers buy something of the catalog? The average gross margin (price - cost) on all products sold through the catalog is 50%.

We can predict the customer behalf with data of our current customers.

2. What data is needed to inform those decisions?

For building up our linear regression model we‘re using the following variable:

1	Average sales amount
2	Customer section

3	Store number
4	Postal code

Step 2: Analysis, Modeling, and Validation

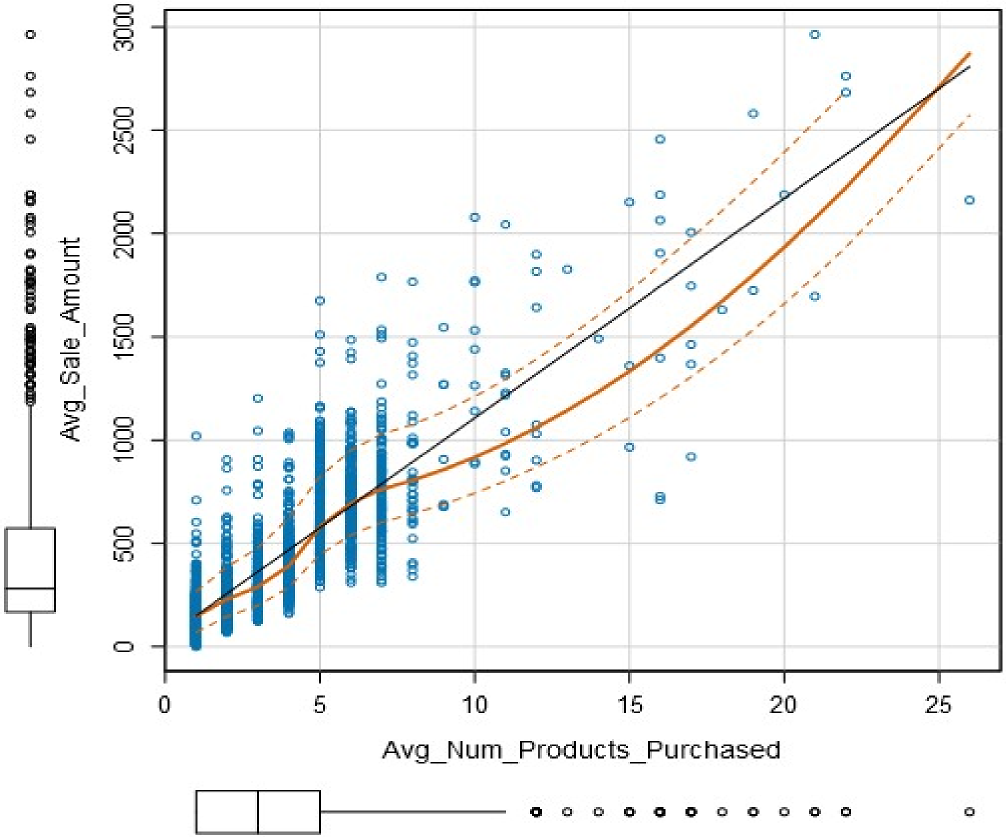
Provide a description of how you set up your linear regression model, what variables you used and why, and the results of the model. Visualizations are encouraged. (500 word limit)

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

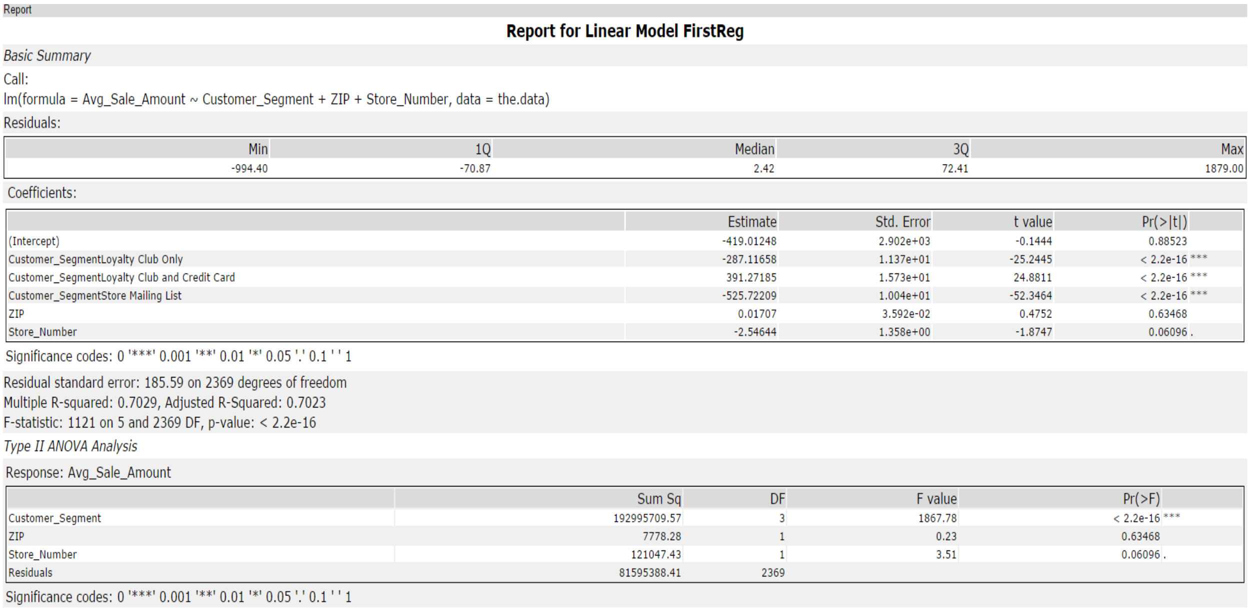
At the minimum, answer these questions:

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 management wants to figure out if we make at least a profit of 10,000$ after sending out the catalogs. So I make sense to use the given data to predict the profit. For predicting the profit we have to use the profit per regular customer. And we can easily check if there is a relationship between our values:



It makes no sense to use for example the postcode or the state for building a linear regression. In Alteryx you're getting this values:



Report

Report for Linear Model FirstReg

Basic Summary

Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + ZIP + Store_Number, data = the.data)

Residuals:

Min	1Q	Median	3Q	Max
-994.40	-70.87	2.42	72.41	1879.00

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-419.01248	2.902e+03	-0.1444	0.88523
Customer_SegmentLoyalty Club Only	-287.11658	1.137e+01	-25.2445	< 2.2e-16 ***
Customer_SegmentLoyalty Club and Credit Card	391.27185	1.573e+01	24.8811	< 2.2e-16 ***
Customer_SegmentStore Mailing List	-525.72209	1.004e+01	-52.3464	< 2.2e-16 ***
ZIP	0.01707	3.592e-02	0.4752	0.63468
Store_Number	-2.54644	1.358e+00	-1.8747	0.06096 .

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

Residual standard error: 185.59 on 2369 degrees of freedom
Multiple R-squared: 0.7029, Adjusted R-Squared: 0.7023
F-statistic: 1121 on 5 and 2369 DF, p-value: < 2.2e-16

Type II ANOVA Analysis

Response: Avg_Sale_Amount

	Sum Sq	DF	F value	Pr(>F)
Customer_Segment	192995709.57	3	1867.78	< 2.2e-16 ***
ZIP	7778.28	1	0.23	0.63468
Store_Number	121047.43	1	3.51	0.06096 .
Residuals	81595388.41	2369		

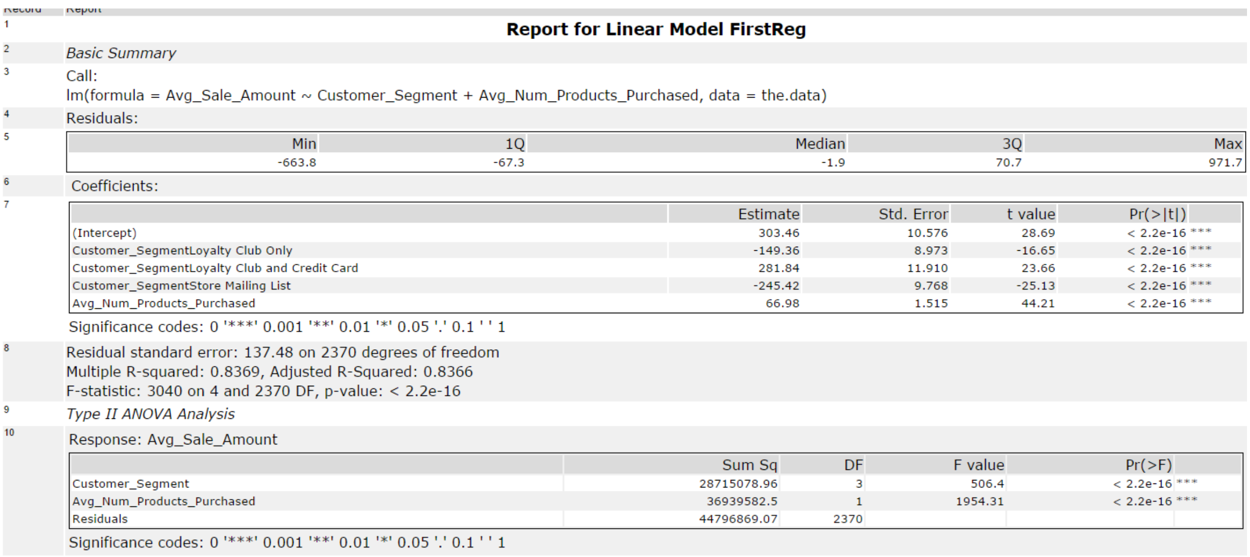
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(figure 02)

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.

*If you check **figure 02** and here inside the row of Pr(>|t|) each value. There are very small and the stars behind the values indicate that the values are quite good. And because of this there is a strong evident, that the customer_segment has impact on the average amount of sales.*

And also you can see for the ZIP Pr(>|t|) value: 0.63 > 0.05 and the store ID Pr(>|t|) value 0.06 > 0.05 is bigger than 5%. So it seems to be, that the ZIP and the sales ID is responsible for the value of the average sales. So for the predict model I'm using the customer segment and the average number of products.



Record Report

1 **Report for Linear Model FirstReg**

2 *Basic Summary*

3 Call:
lm(formula = Avg_Sale_Amount ~ Customer_Segment + Avg_Num_Products_Purchased, data = the.data)

4 Residuals:

5

Min	1Q	Median	3Q	Max
-663.8	-67.3	-1.9	70.7	971.7

6 Coefficients:

7

	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
F-statistic: 3040 on 4 and 2370 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

(figure 03)

*In **figure 03** you can see that the r^2 value is 0.8366 so approx 84%. So 84% of the average sales per customer is explained by the model. The F-statistic is 99% this means all of the used variables having effect on the average sales per customer.*

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.

Based on figure 03 we' re getting the following regression equation:

Y= 303.46 + 0 * Customer_Segment_Credit_Card_Only – 149.36 * Customer_Segment_Loyalty_Club_Only + 281.84 * Customer_Segment_Loyalty_Club_And_Credit_Card – 245.42 * Customer_Segment_Mailing_List + 66.98 * Average_Number_Products_Purchased

*0 * Customer_Segment_Credit_Card_Only is eliminating the credit_card_only customers.*

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?

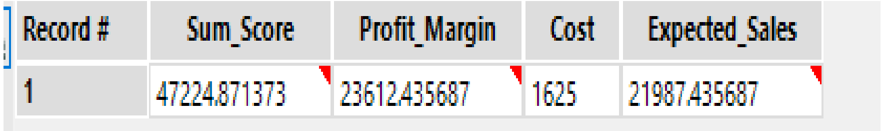
2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process)

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

Based on our modell we‘re getting the following formula:

*Expected_Score = Score * Score_Yes*
*Profit_Margin = Sum_of_Expected_Score * 0.50*
*Cost = 250 * 6.50$*
Expected_Sales = Profit_Margin – Cost

We’re getting the following values:



Record #	Sum_Score	Profit_Margin	Cost	Expected_Sales
1	47224.871373	23612.435687	1625	21987.435687

So the expected margin is: $23,612.43, the cost: $1,625 and without the cost we‘ re getting the expected sales: $21,987.43.

$21,987.43$ > $10,000 so we recommend the management to sent out the catalog to the 250 new 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.