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?
   We need to decide whether we should send out catalogs to our 250 new customers. Our profit must exceed $10,000 in order for this effort to be worthwhile.

2. What data is needed to inform those decisions?
   We need existing customer data, sales numbers, and whether a customer is likely to respond to a catalog.

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.

   I used a correlation matrix to find which continuous variable would be the best predictor. A correlation of 1 indicates a good predictor. I also looked at the corresponding p-values to make sure a relationship exists between the two variables.

   Average Number of Products Purchased is a good predictor, while the Number of Years as a Customer is not.

   Customer ID, Zip, Store Number are not good predictors and are non-continuous values.

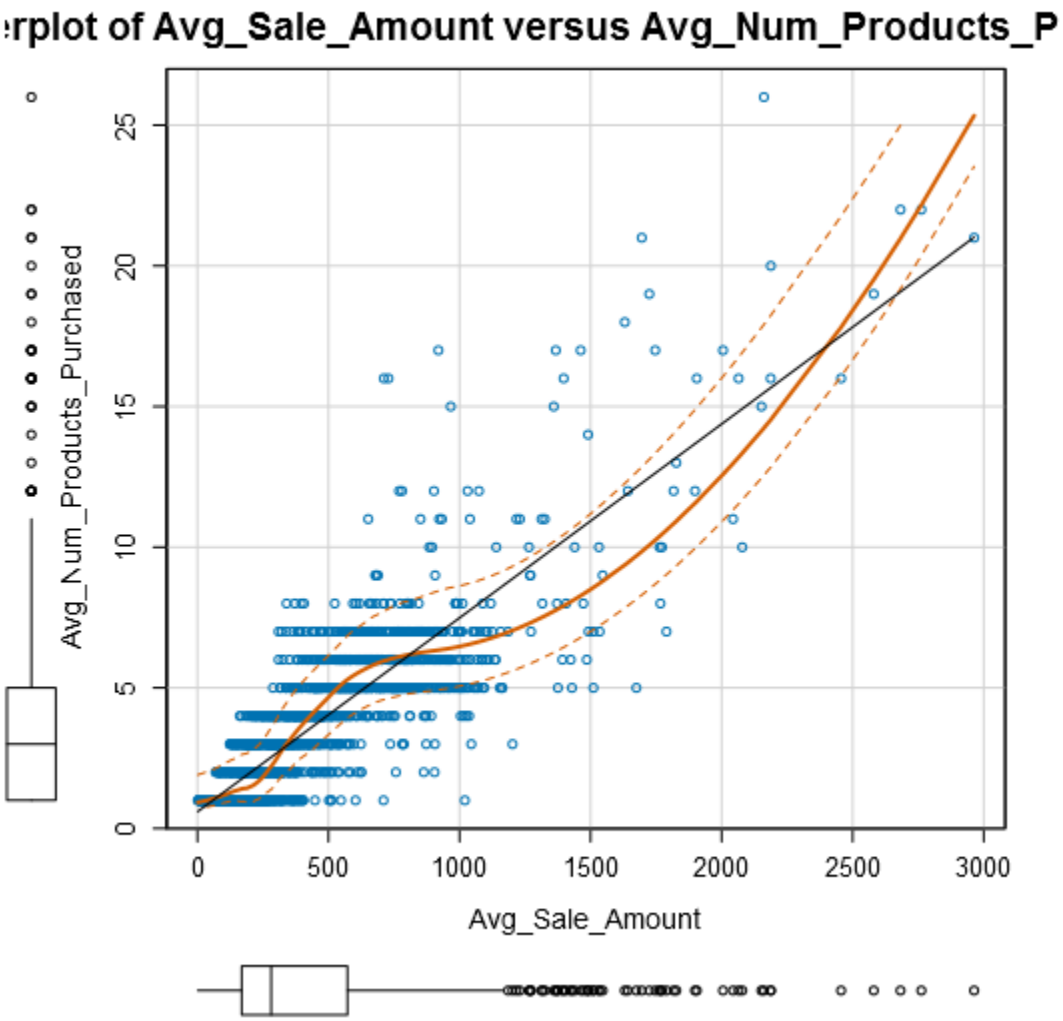
Full Correlation Matrix

	Customer.ID	ZIP	Avg.Sale.Amount	Store.Number	Avg.Num.Products.Purchased	X..Years.as.Customer
Customer.ID	1.0000000	0.0021590	0.0382352	-0.0233227	0.0601359	0.0151644
ZIP	0.0021590	1.0000000	0.0079728	-0.1489063	0.0017896	0.0016432
Avg.Sale.Amount	0.0382352	0.0079728	1.0000000	-0.0079457	0.8557542	0.0297819
Store.Number	-0.0233227	-0.1489063	-0.0079457	1.0000000	-0.0115250	-0.0095729
Avg.Num.Products.Purchased	0.0601359	0.0017896	0.8557542	-0.0115250	1.0000000	0.0433464
X..Years.as.Customer	0.0151644	0.0016432	0.0297819	-0.0095729	0.0433464	1.0000000

Matrix of Corresponding p-values

	Customer.ID	ZIP	Avg.Sale.Amount	Store.Number	Avg.Num.Products.Purchased	X..Years.as.Customer
Customer.ID		9.1625e-01	6.2455e-02	2.5589e-01	3.3703e-03	4.6010e-01
ZIP	9.1625e-01		6.9776e-01	3.0154e-13	9.3054e-01	9.3621e-01
Avg.Sale.Amount	6.2455e-02	6.9776e-01		6.9873e-01	0.0000e+00	1.4679e-01
Store.Number	2.5589e-01	3.0154e-13	6.9873e-01		5.7454e-01	6.4101e-01
Avg.Num.Products.Purchased	3.3703e-03	9.3054e-01	0.0000e+00	5.7454e-01		3.4659e-02
X..Years.as.Customer	4.6010e-01	9.3621e-01	1.4679e-01	6.4101e-01	3.4659e-02	

Here is a scatter plot of avg sale amount and the average number of products purchased to illustrate the correlation.



I examined the pvalues of the other possible predictors in the linear regression model. Customer Segment and Avg Number of Products Purchased are both good candidates based on the p value of 2.2e-16

Response: Avg.Sale.Amount

	Sum Sq	DF	F value	Pr(>F)
Customer.Segment	27368293.94	3	481.89	< 2.2e-16 ***
ZIP	1407752.36	85	0.87	0.78509
Store.Number	203608.81	9	1.2	0.2935
Responded.to.Last.Catalog	115725.89	1	6.11	0.01349 *
Avg.Num.Products.Purchased	35331983.8	1	1866.32	< 2.2e-16 ***
X..Years.as.Customer	75730.43	1	4	0.04561 *
Residuals	43049952.34	2274		

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

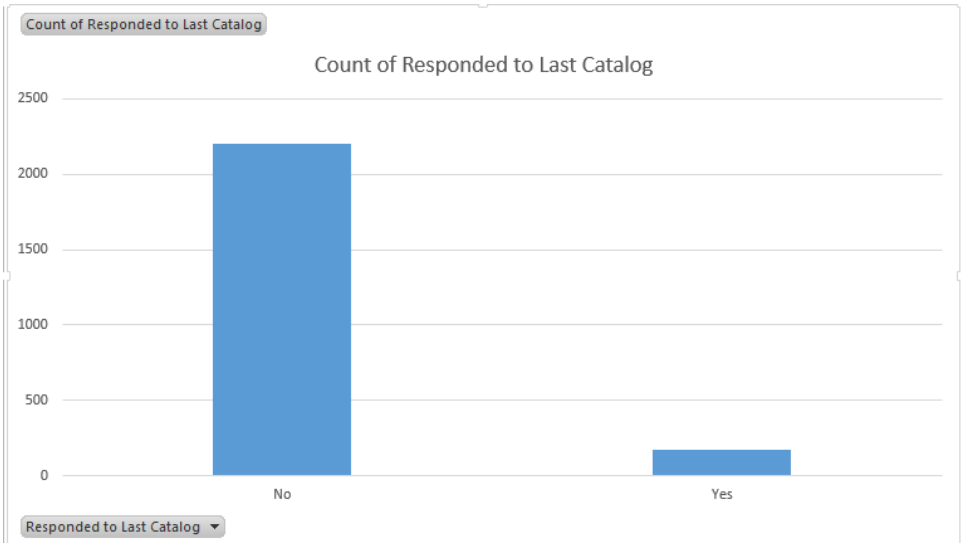
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.501e+02	31.267	1.120e+01	< 2.2e-16 ***
Customer.SegmentLoyalty Club Only	-1.484e+02	9.150	-1.622e+01	< 2.2e-16 ***
Customer.SegmentLoyalty Club and Credit Card	2.873e+02	12.210	2.353e+01	< 2.2e-16 ***
Customer.SegmentStore Mailing List	-2.411e+02	10.030	-2.403e+01	< 2.2e-16 ***
ZIP80003	-3.304e+01	31.149	-1.061e+00	0.28893
ZIP80004	-2.537e+01	29.224	-8.682e-01	0.38539
ZIP80005	-1.948e+01	29.902	-6.516e-01	0.5147
ZIP80007	-2.152e+01	57.629	-3.735e-01	0.70882

…

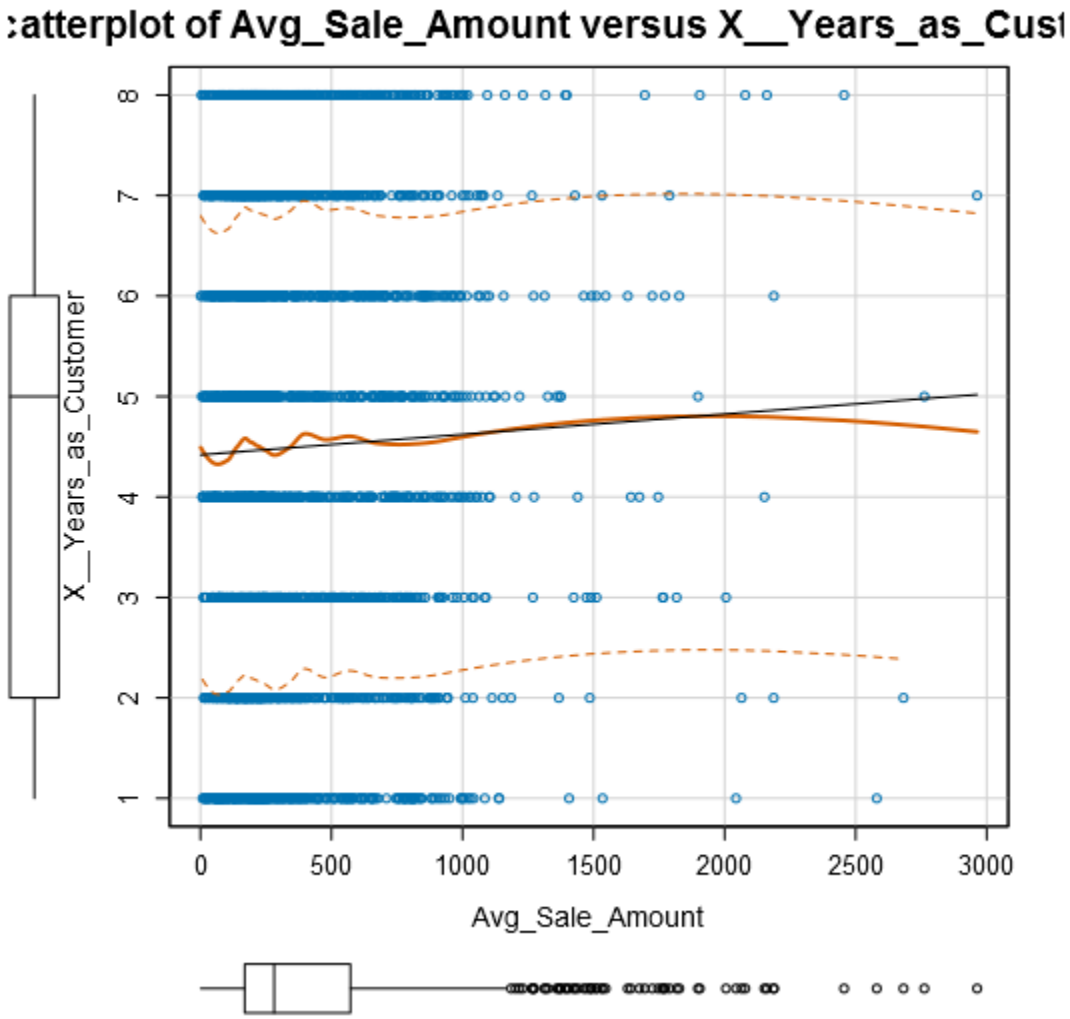
ZIP80602	-7.339e+01	73.751	-9.951e-01	0.31977
ZIP80640	-2.913e+02	140.461	-2.074e+00	0.03818 *
Store.Number101	-5.394e-03	14.434	-3.737e-04	0.9997
Store.Number102	1.530e+01	23.144	6.609e-01	0.50875
Store.Number103	2.414e+00	18.470	1.307e-01	0.89602
Store.Number104	-1.204e+01	14.495	-8.303e-01	0.40643
Store.Number105	-1.322e+01	12.464	-1.061e+00	0.28884
Store.Number106	-2.030e+01	14.675	-1.384e+00	0.16665
Store.Number107	-2.496e+01	15.943	-1.566e+00	0.11754
Store.Number108	-1.532e+01	17.792	-8.609e-01	0.38938
Store.Number109	1.147e+01	21.318	5.380e-01	0.59061
Responded.to.Last.CatalogYes	-2.852e+01	11.535	-2.472e+00	0.01349 *
Avg.Num.Products.Purchased	6.690e+01	1.549	4.320e+01	< 2.2e-16 ***
X..Years.as.Customer	-2.513e+00	1.256	-2.000e+00	0.04561 *

All non-valuable predictors were removed from the model.

The data was also sliced to see if users responded to the catalog historically. Most users did not.



As more evidence, this scatterplot shows that there is no correlation between the average sale amount and the number of years as a customer.



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 P values are very close to zero, 2.2e-16, indicating that the observed results did not occur by chance and that a relationship exists between the predictor and target.

   The Multiple R^2 value is .8369 and the Adjusted R^2 value is .8366. An R^2 value .75 or higher is a decent indicator of the model.

	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 DF, p-value: < 2.2e-16

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)

   = 303.46 – (149.36 * Loyalty Club Segment) + (281.84 * Loyalty Club and Credit Card Segment) – (245.42 * Mailing List Segment) + (0* Credit Card Segment) + (66.98 * Avg Number of Products Purchased).

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?

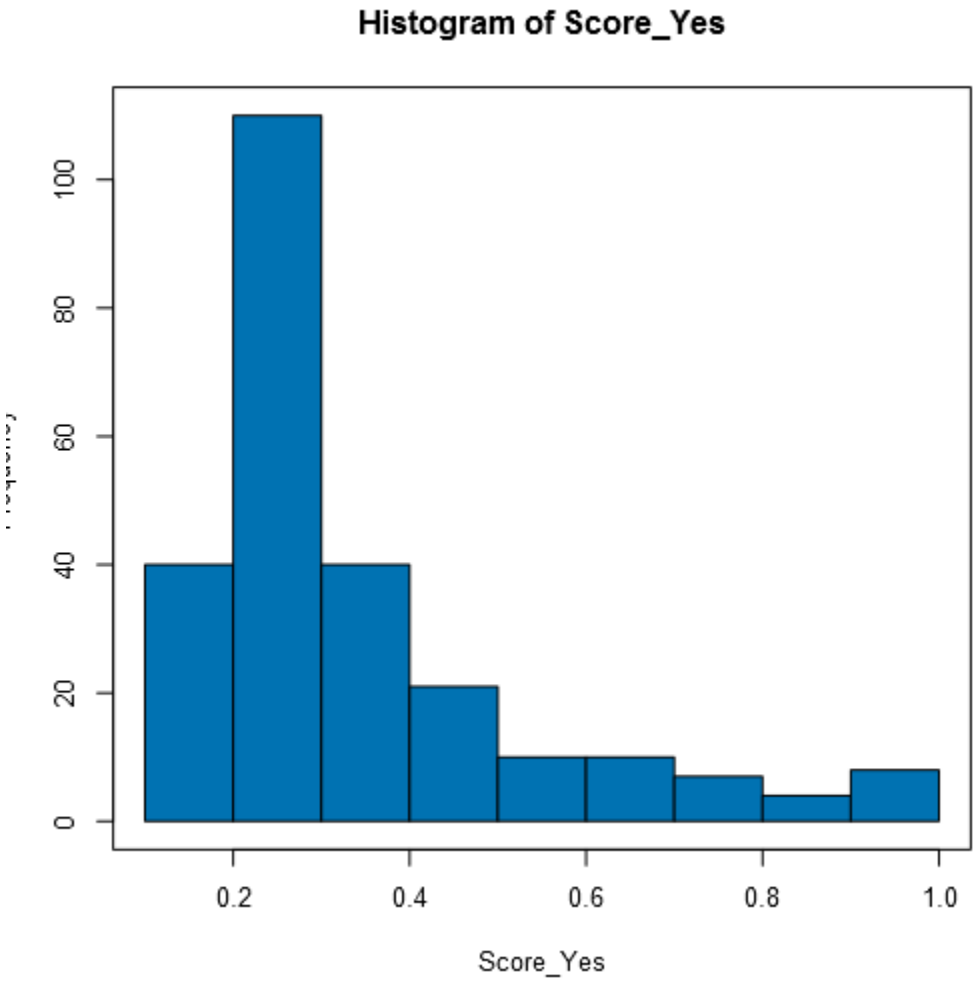
   I would recommend sending catalogs to the 250 new customers.

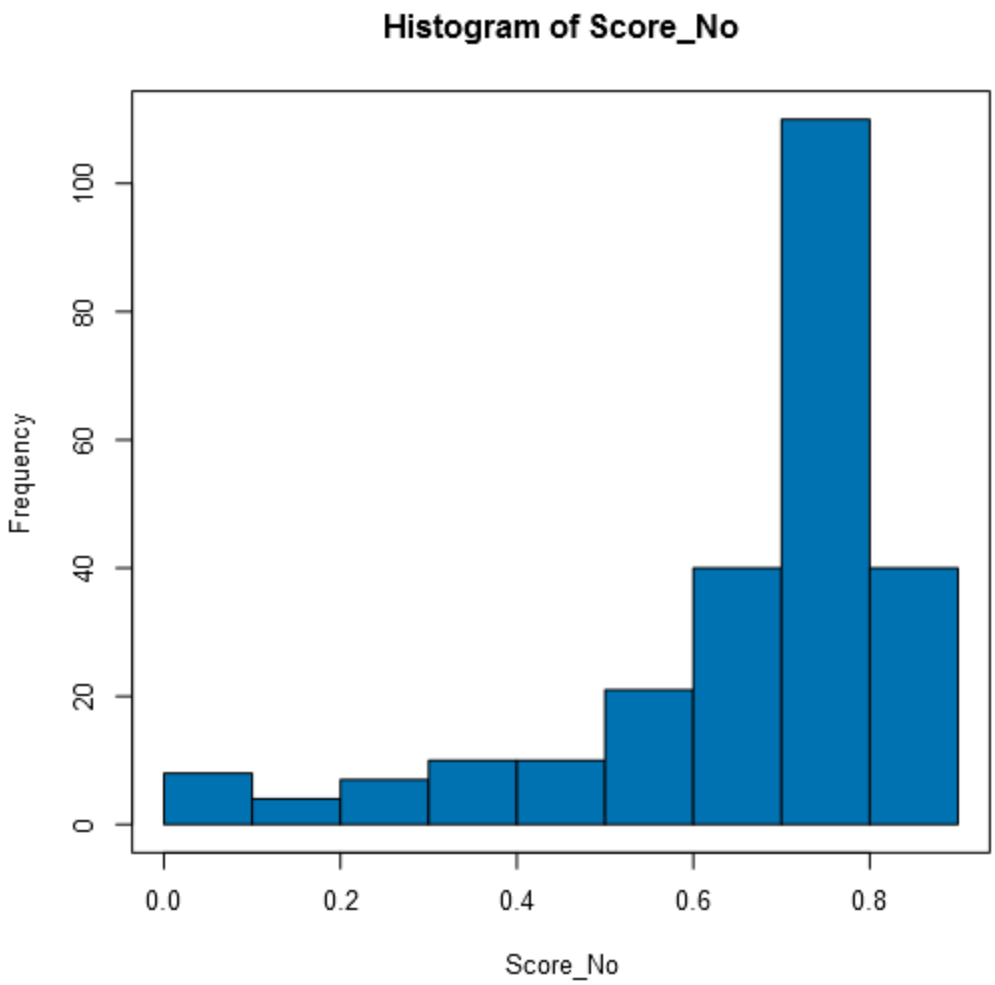
   The expected profit would be $21,987.44, based on whether customers would respond to the catalog, minus margin and the cost of production, see below formula.

   Profit = (sum(avg sales amount * score) * gross margin) –
   (cost of catalog production * number of flyers created and sent)

   The catalog response rate has a higher and stronger probability that customers **will not respond** to the catalog (No_Score values) than they **would respond** to the catalog (Yes_Score values) values.

   However, because $21,987.44 exceeds the target of $10,000. Our catalog mailing efforts would be profitable enough.


Histogram of Score_Yes
Frequency
0
20
40
60
80
100
0.2
0.4
0.6
0.8
1.0
Score_Yes



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

   I used a multiple linear regression model using customer segment and avg number of products purchased as predictors for average sale amount. Other columns were evaluated, but the p-variable was too high and the impact to the R^2 value negligible. They were removed from the model.

   I examined the probability of a customer purchasing from catalog using a histogram for the Yes/No Score.

Profit = (sum(avg sales amount * score) * gross margin) –
(cost of catalog production * number of flyers created and sent)

Profit, w/ catalog = (sum(avg sales amount * yes_score) *.5) – (6.50 * 250)

Profit, w/out catalog = (sum(avg sales amount * no_score) *.5) – (6.50 * 0)

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

   The expected profit would be $21987.44

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.