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? In this project I need to predict the sales for a 250 new customers that compose a mailing list. The idea here is to predict if the sum of these future sales will be beyond the $10,000 mark, and otherwise, no mailing actions will be performed by the company.
2. What data is needed to inform those decisions? I will need historical data regarding customers from the company and their associated sales number (training set), such as the information about the new customers (test set).

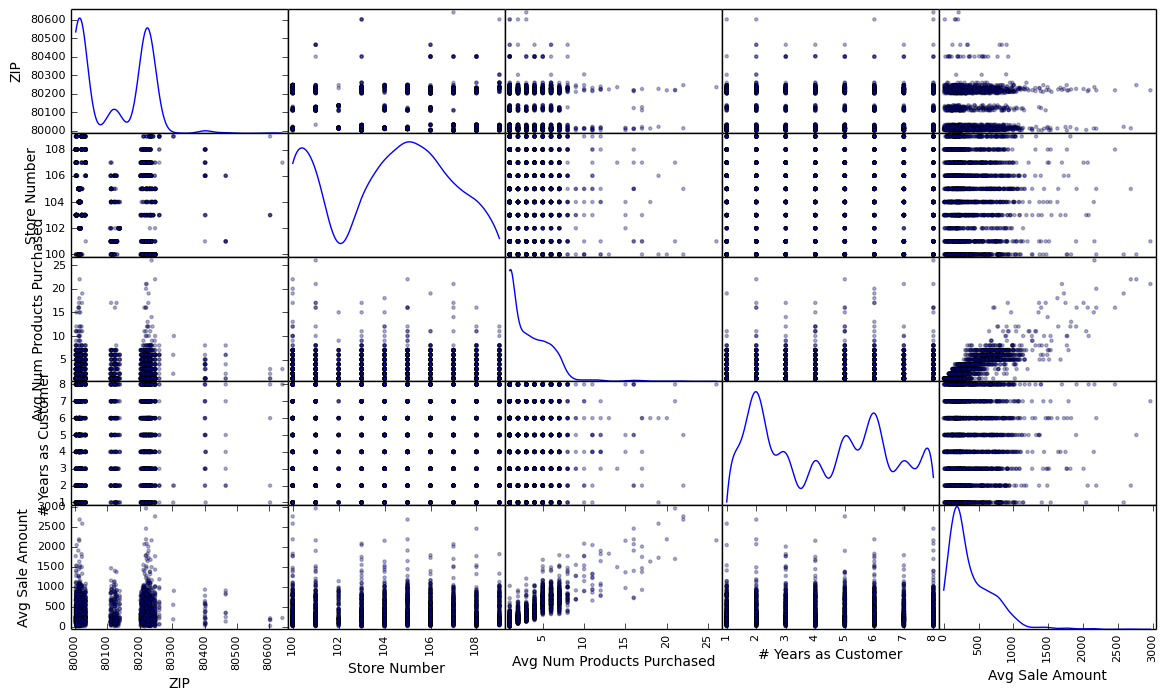
## Step 2: Analysis, Modeling, and Validation

*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)](https://classroom.udacity.com/courses/ud976/lessons/4e33b70a-72a4-47cb-959a-28632ae6aaff/concepts/631d190c-8626-4dd7-92df-f5bd96913c48) 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](https://classroom.udacity.com/nanodegrees/nd008/parts/c0b53068-1239-4f01-82bf-24886872f48e/modules/bf705147-0d7c-4492-842a-698a6410a8a3/lessons/4e33b70a-72a4-47cb-959a-28632ae6aaff/concepts/631d190c-8626-4dd7-92df-f5bd96913c48) to help you explore your data and use scatterplots to search for linear relationships. You must include scatterplots in your answer. I’ve used Pandas (see the Jupyter notebook in this repo) to work on this project, and before anything, I’ve converted all categorical data into dummy variables so they would be processed by the Linear Regression algorithm. I then used pandas to plot the scatter matrix to each the linear correlation amongst each pair of variables. The following plot shows this matrix. I can’t see any obvious correlations from these plots, but there seems to be some kind of correlation between ‘Avg Num Products Purchased’ and ‘Avg Sale Amount’ (which is somehow expected, since if you purchase more products the same amount is also expected to increase). To be sure I wasn’t missing anything, I also plotted a heatmap with the correlations between each pair of attributes, which is also presented below. On the heatmap, we can confirm this correlation between these two features, which has a pearson correlation of 0.855754 to be precise. A possible question here would be why the correlation doesn’t seem so clear in the scatter matrix, and my argument is that we have many outliers that don’t allow us to clearly see the “dense” part of the plot (lower left) which has a lot of data. Given that, I will try to learn a model using the ZIP, Store Number, Avg Num Products Purchased, #Years as Customer to predict the Avg Sale Amount for each customer. (These plots don’t show the correlations between the dummy variables, because it would jeopardize visualization!)





1. 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 we analyze the R squared result obtained by my linear regression model, we’ll see that it resulted in a value of 0.8309, which depicts a high determination coefficient. In the next question I present the regression equation, which uses only attributes with a p-value below 0.05. The p-value for each variable that I originally had in my dataset is:

* ZIP has a p-value of 0.697757997163
* Store Number has a p-value of 0.698733999956
* Avg Num Products Purchased has a p-value of 0.0
* # Years as Customer has a p-value of 0.146794828448
* ['Customer Segment']\_Credit Card Only has a p-value of 1.59755756527e-105
* ['Customer Segment']\_Loyalty Club Only has a p-value of 0.779551878195
* ['Customer Segment']\_Loyalty Club and Credit Card has a p-value of 3.76987503219e-224
* ['Customer Segment']\_Store Mailing List has a p-value of 3.33692101865e-305
* ['City']\_Arvada has a p-value of 0.504135266812
* ['City']\_Aurora has a p-value of 0.7102767688
* ['City']\_Broomfield has a p-value of 0.832118580649
* ['City']\_Castle Pines has a p-value of 0.388017931143
* ['City']\_Centennial has a p-value of 0.4678272136
* ['City']\_Commerce City has a p-value of 0.00833292916966
* ['City']\_Denver has a p-value of 0.636015065553
* ['City']\_Edgewater has a p-value of 0.518910326034
* ['City']\_Englewood has a p-value of 0.601004098726
* ['City']\_Golden has a p-value of 0.734093555445
* ['City']\_Greenwood Village has a p-value of 0.416066344277
* ['City']\_Highlands Ranch has a p-value of 0.900987857771
* ['City']\_Lakewood has a p-value of 0.414479682012
* ['City']\_Littleton has a p-value of 0.553152903339
* ['City']\_Louisville has a p-value of 0.762756806876
* ['City']\_Northglenn has a p-value of 0.533523534953
* ['City']\_Parker has a p-value of 0.319914199047
* ['City']\_Thornton has a p-value of 0.923395421615
* ['City']\_Westminster has a p-value of 0.906476420085
* ['City']\_Wheat Ridge has a p-value of 0.509521078639

1. 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)

Y = 67.01 \* Avg Num Products Purchased + 149.35 \* ['Customer Segment']\_Credit Card Only + 431.40 \* ['Customer Segment']\_Loyalty Club and Credit Card + -96.03 \* ['Customer Segment']\_Store Mailing List + -28.36 \* ['City']\_Commerce City

**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? Yes, given my computations, the profit would be above the threshold that the manager has set, which was of USD 10,000.
2. How did you come up with your recommendation? (Please explain your process so reviewers can give you feedback on your process) I learned a linear regression model from historical sales to predict the values of sales for 250 future customers, which resulted in a value of USD 69137.76. Assuming that only 30% of the customers would actually purchase the products in the catalog, the revenue would be of USD 20741.33. Evidently, there is a cost to print and distribute all the 250 catalogs, which is 250 \* 6.50 = USD 1625.00, which should be then decreased from the revenue.
3. What is the expected profit from the new catalog (assuming the catalog is sent to these 250 customers)? Given the latter process, if we take the revenue (USD 20741.33) and decrease the printing and distribution costs (USD 1625.00), the profit would be of USD 19116.33, which is above the threshold set by the manager (USD 10000), and thus, my recommendation would be of printing and distributing the catalog.

Before you Submit

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/186/view) here. Reviewers will use this rubric to grade your project.