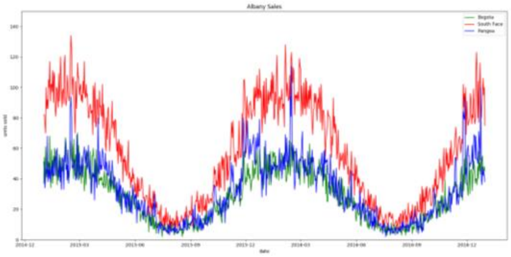
Exploratory Analysis
IES Data Science Challenge
Author: Kan Ito
All extraction, manipulation, and analysis executed in Python.

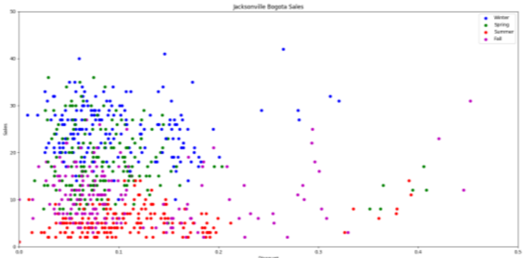
7/12/2017: All data was extracted simply. First two sets of files are categorical data that defines the boundaries of the data in the 3rd and 4th files. Empty cells of data were observed for the 3rd file. Those empty prices were filled with expected prices based on retail price and the national discount program. The 4th file is large. However, we are interested in the sales so the data from the transactional data was converted to sums instead of saving each row of data.

7/13/2017: Data visuals are complete. Visuals of the first two data sets do not reveal anything interesting so far. Third and fourth files are visualized for sales over a period of time. This visualization reveals in sales for time of year. Furthermore, visualizing among different brands and locations reveal variation in sales. It is of interest to look into research of :

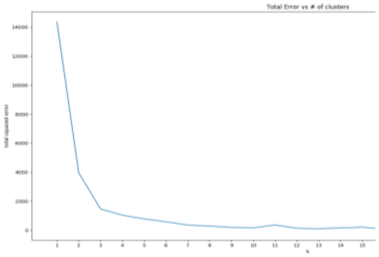
- Jacket sales and geography as a factor in retail sales
- Correlations between retail discounts and impact on profits
- Consumer behavior with brands' and public consumption habits

7/14/2017: From more plotting and visuals, it seems seasons are the largest influence on sales. To understand the behavior of sales we would like to form some predictive modeling technique to eventually provide recommendations. Although seasons are a large factor in the sales, we are also interested in other variables including: date of year, seasonal discount, holiday discount, store-basis discount, brand, location. The three types of discount is assumed not observed by the consumer and will be forged into one category of discount that encompasses seasonal discount, holiday discount, and store-basis discount.

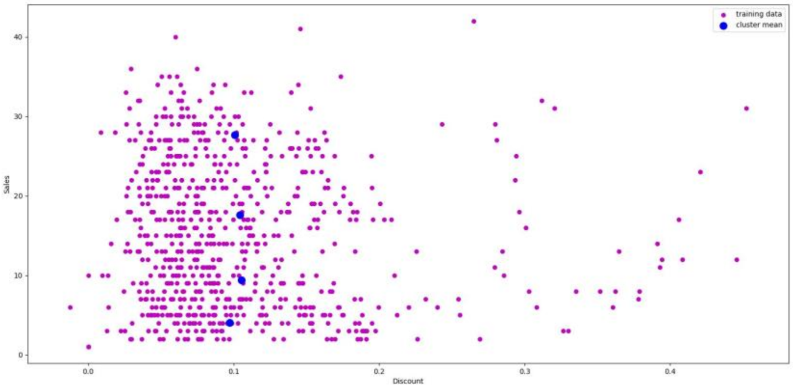




7/16/2017: Clustering is a common method utilized in modeling sales. The model revealed a clustering of 3 or more means based on the data. The error is significantly reduced after k = 3 or 4 means.



The clustering model results are shown below showing the means points (4 total).



However, this reveals means points that are basically in a straight line hovering around discount = 0.2 (20%) which indicates low correspondence in discount with sales and more importantly, this model does not take advantage of the other known factors of the data such as time of year, brand, and location.
7/17/2017: Multiple regression is attempted on one brand at one location. The independent variables attempted are discount, season, and constant. Then, several more variables including products of variables were added into the regression. Note that seasons were generated with discretized values among 365/2~183 days that signaled warmer and colder days of the year.

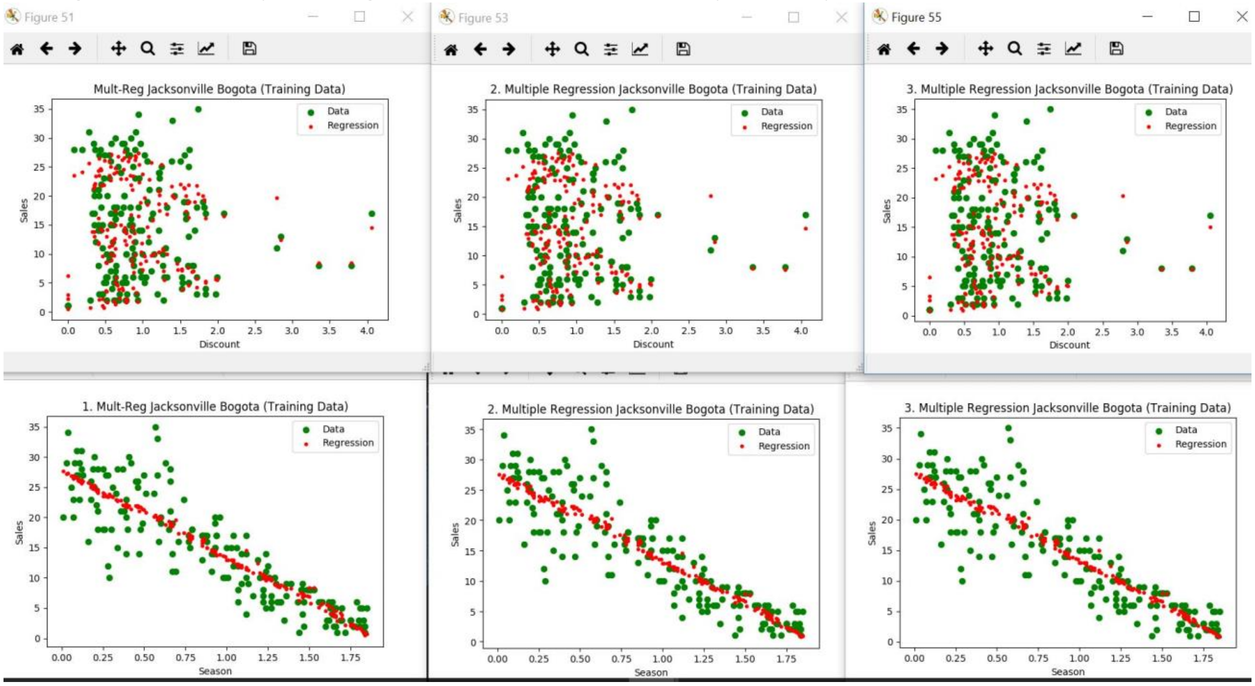
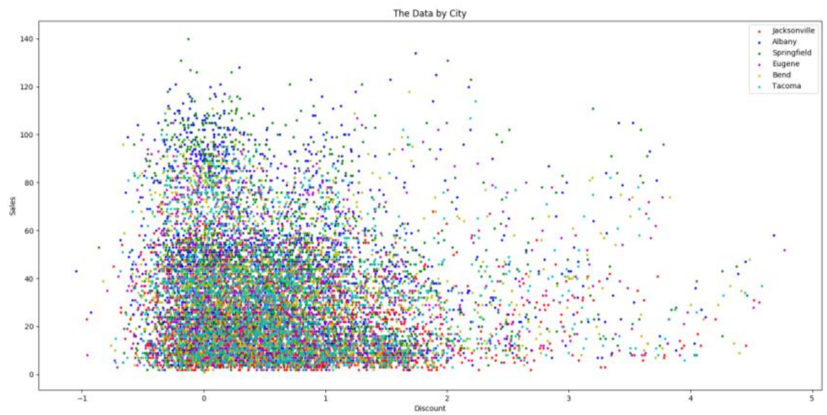
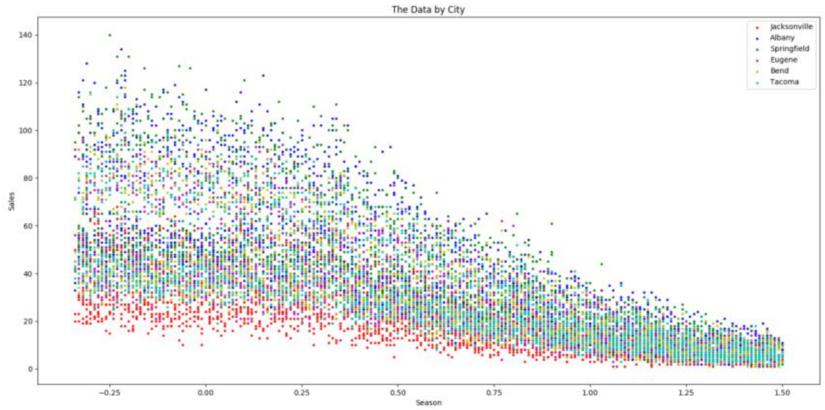


Figure 51
Figure 53
Figure 55
Mult-Reg Jacksonville Bogota (Training Data)
2. Multiple Regression Jacksonville Bogota (Training Data)
3. Multiple Regression Jacksonville Bogota (Training Data)
1. Mult-Reg Jacksonville Bogota (Training Data)
2. Multiple Regression Jacksonville Bogota (Training Data)
3. Multiple Regression Jacksonville Bogota (Training Data)
Data
Regression
Sales
Discount
Season
0
5
10
15
20
25
30
35
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

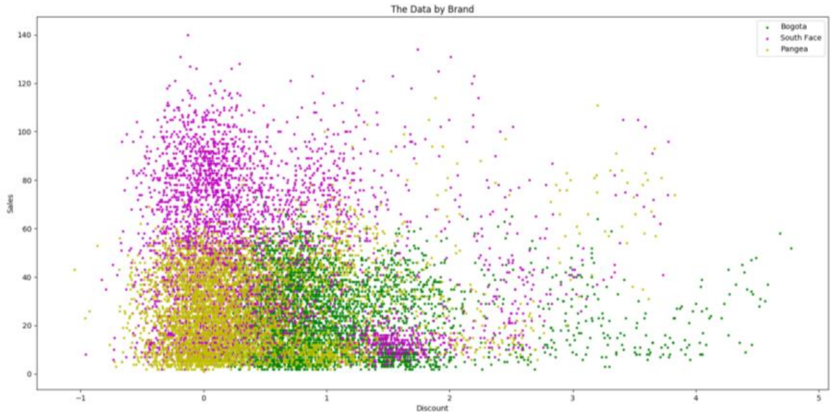
Note that there is very little difference among the three different multiple regressions. The extra coefficients in the more complicated regressions are essentially insignificant. R-squared values showed strength just below 0.8 with not much variation among the different models. The multiple regression is based on gradient descent minimization.

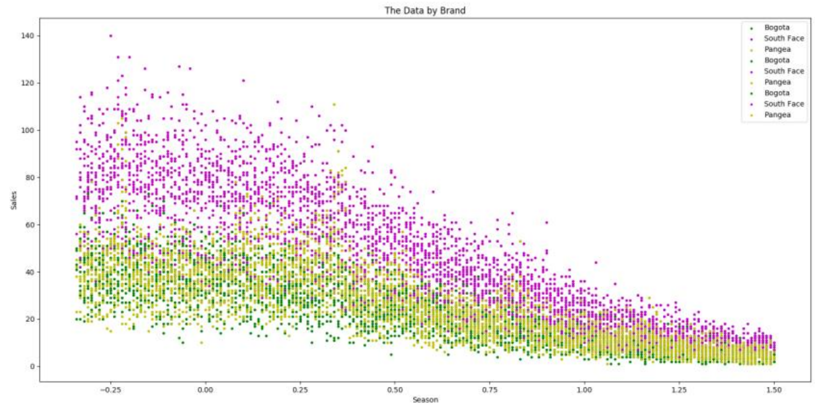
7/21/2017: By investigating the bands of sales data by brand and/or city categories, the sales showed definitive separation. We can easily see the bands when plotted for the season. When plotted against discounts, the separation was harder to see visually.



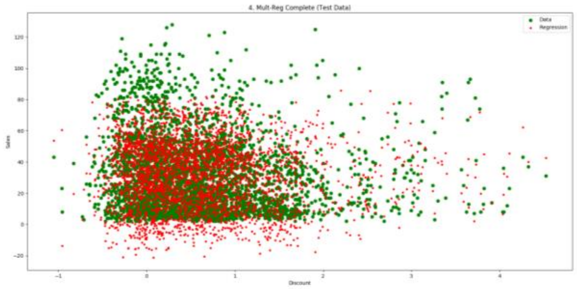


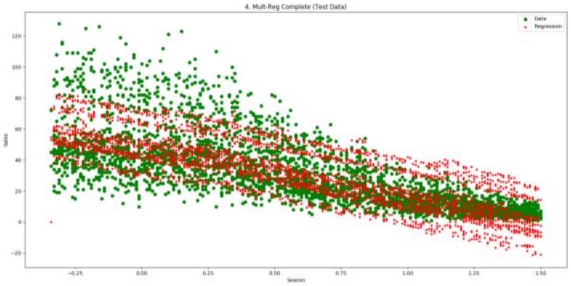
Similarly, there was reason to believe there exists an effect on sales due to brand as well.





Accordingly, multi-regression was run with binary coefficients for all cities and brands.





The following variables were used for the regression model: discount, constant, season, all cities, all brands, season*discount, discount squared, season squared. Unsurprisingly, R-squared values remain around 0.8

7/22/2017: Since we are interested most in utilizing discounts to improve sales, discount v sales behavior was further investigated. Furthermore, it was of interest where specifically the discount program can be more effective. From the multiple regression model, it was found that sales were more volatile to discounts during winter seasons. By executing two separate linear regressions for two sides of the season spectrum, it was found that the linear coefficient for winter days were higher; discount programs reveal larger effects on sales during the winter. Naturally, that became a recommendation. Similarly, the regressions were conducted for each city and brand revealing preferable cities and brands to utilize discounts more than others. These became additional recommendations.

7/23/2017: It is of further interest to investigate correlations among different combinations of cities and brands. Additionally, the discounts can be unbundled to reveal strengths of the discount programs where we can provide recommendations more specifically on the magnitude of discounts and perhaps perform a linear optimization defining profit as the maximizer with constraints defined by prices, minimum store/brand requirements, etc. For the purpose of recommendations proposed by this report, it is not of prime importance to conduct standard errors of regression coefficients and perhaps a more robust ridge regression technique. Moving forward, it should be done to validate and form a stronger argument for the recommendations. Note that all modeling was performed on training data (about 25-30% of the data set) and tested on a randomized sample for verification.