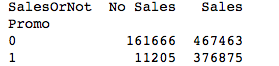
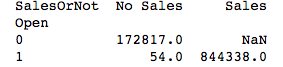
**Milestone3 Summary**

In the EDA process, we need to learn more about the data. A general description shows that the training data set has 8 variables, and 6 of them (Store, DayOfWeek, Open, Promo, and StateHoliday, SchoolHoliday) can be used as predictors for modeling the Sales. There is missing data in the Open column so we need to handle this later. There are 1115 stores in the training data and 856 stores in the testing data, and all the stores in the testing data is in the training data set.

**Research on train.csv**

Holidays vs. Sales: Next we want to study the relationship between the variables and the sales. The boxplot of holiday versus sales shows that there is an increase in sales during state holidays (See Figure1), while the sales amount remain roughly the same whether it’s school holiday or not. It’s reasonable since most customers are adults and they don’t have time to go shopping during school holidays (See Figure2).

Promotion vs. Sales: After that we want to check the effects of promotion on the sales. The boxplot of sales grouped by promotion shows that the promotion will boost the sales a little bit on average (See Figure3), while the distribution still looks the same. So the promotion does have impact on sales, but they are not as important as state holidays. The following contingency table also shows that there will most like be sales if the store is having a promotion. The contingency table of open versus sales shows that if the store is closed, there won’t be any sales, which is reasonable and should be considered in our model.

Power of Sundays: The analysis on weekdays versus sales shows that people tends to do shopping on Sundays (See Figure 4), followed by Monday and Friday. It’s interesting to see that people do less shopping on Saturday, although they don’t need to work on Saturday. Now that we know Sunday is very important to sales, we want to learn more about it. The time series plots on Store 82 and Store 85 show that being Sunday will increase the sales in Store 85 (See Figure5), but the sales on Sunday in Store 82 is 0 (See Figure6). The reason for this might be that Store 82 is closed on weekends, so we need to be aware of the interaction between the variables.

Seasonal Impact: Next we want to check whether there are seasonal effects on the sales. The time series plot (See Figure 7) over year shows that people tends to buy more stuff during the holiday seasons (November and December), which agrees with our common belief.

**Research on train.csv and store.csv**

Competition Distance vs. Sales: Next we look into the store data. There are 9 variables in this data set, which can help us tell which stores are similar to each other, or directly help us to predict the sales. The scatter plot of competition distance after taking log and the log(sales) shows a slightly negative relationship, which indicates that the longer the competition distance is, the less the sales will be.(Figure 9)

Promotion Interval and Sales: Next we move to the study of promotion interval and sales. The boxplot of sales in different promotion interval group are roughly the same. However, if we can divide them to months, there might be some months that promotion will be more impactful. (See Figure 10)

Store Types Research: Then we check the impact of different store types on sales. The time series plot shows that store type B will have huge impact on sales (See Figure 11, Figure 12), while the seasonal effect on different store types are roughly the same.

Assortment Research: The time series plot of sales for different assortment also shows that assortment B will boost the sales a lot (See Figure 13, Figure 14). However, the previous contingency table shows that there are only a few stores are store type B or assortment type B, so this observation may not be very useful in the modeling stage.

**Dealing with Missing Data and Filling Data Gaps**

In training dataset, we see the number of stores reported every day is not exactly the same (see Figure 15). There are 180 stores with missing values from 7/31 2014 to 12/31 2014. To fill the gap of these 180 stores, we take a couple of methods to do data cleaning and wrangling:

1. Taking one store as an example to learn time series pattern of the dataset. (See Fig16)
2. For Sales column, I take the median of the Sales of the same store from the previous 5 Sales data points of the same weekday to fill it in.
3. For Promos column, taking the majority voting of Promos of other stores at the same day.
4. For Open column, taking the majority voting of Open of other stores at the same day.
5. For School Holiday, taking the majority voting of School holiday of other stores at the same day.

Figure1:

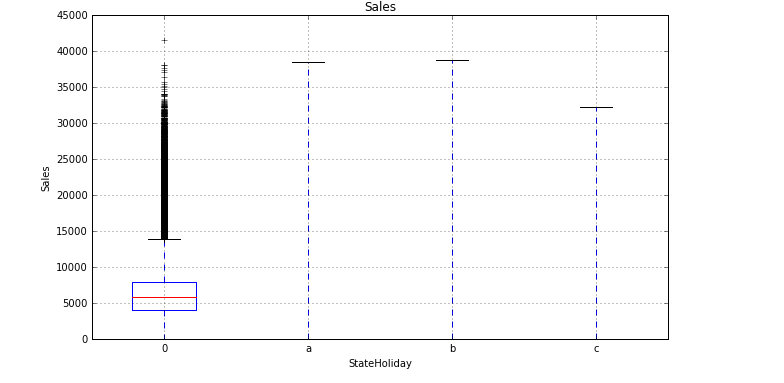


Figure2:

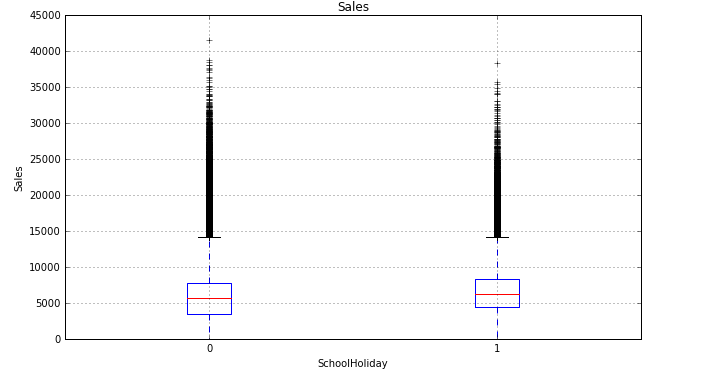


Figure3:

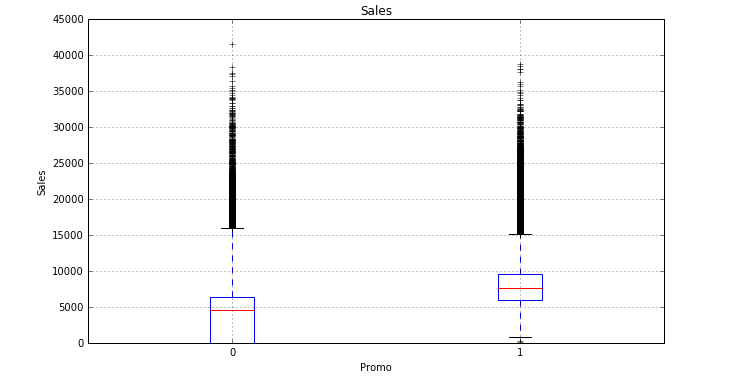


Figure4:

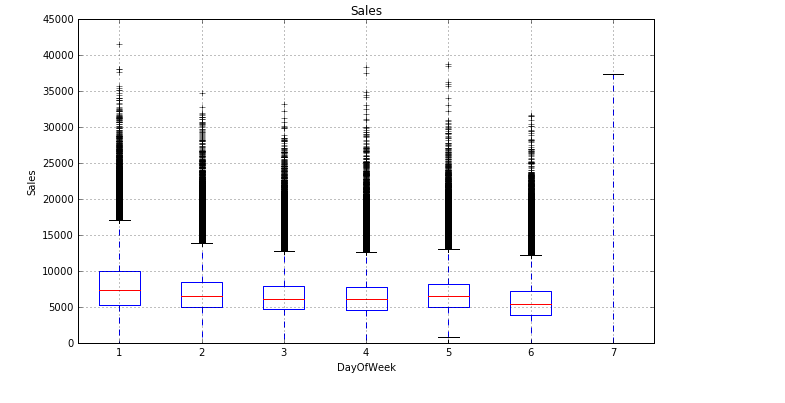


Figure5:

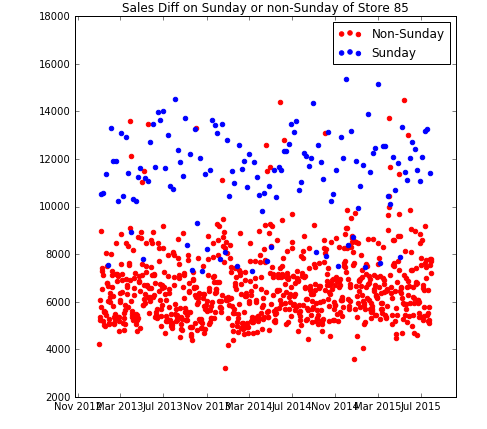


Figure6:

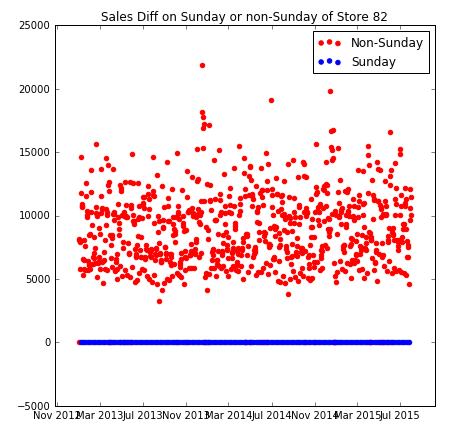


Figure7:

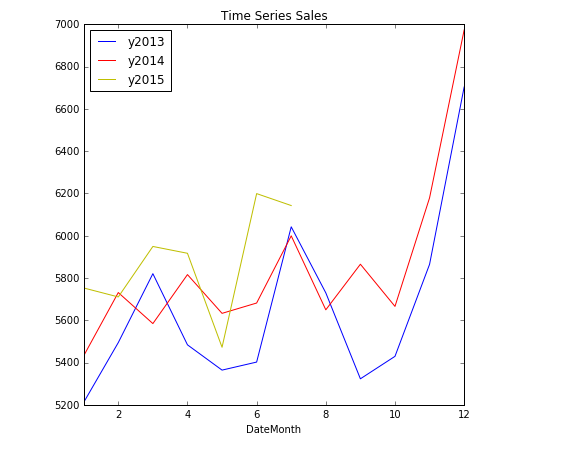


Figure8:

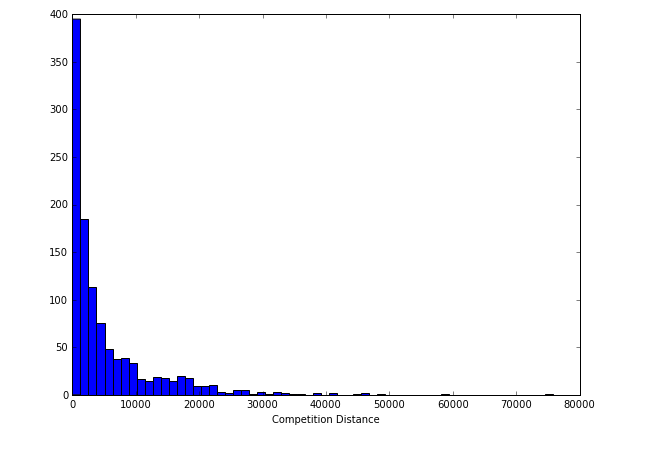


Figure9:

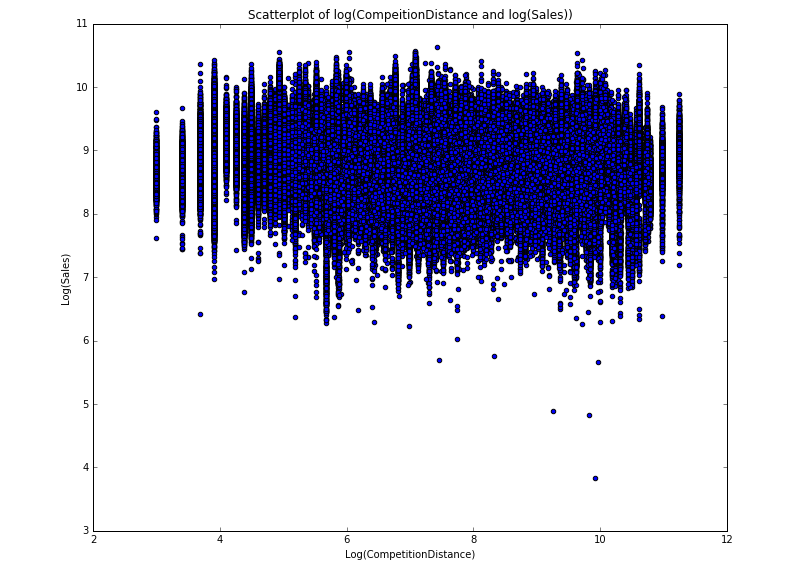


Figure10:

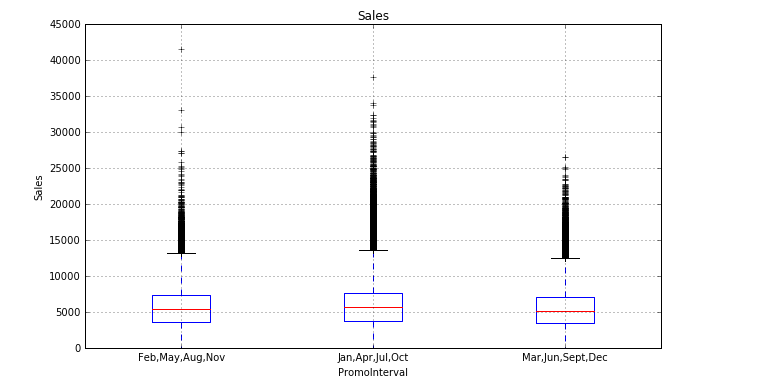


Figure11:

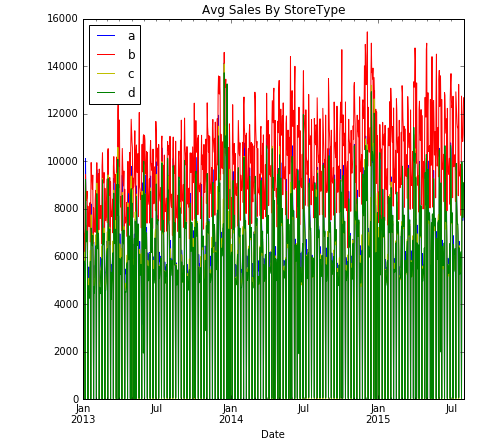


Figure12:

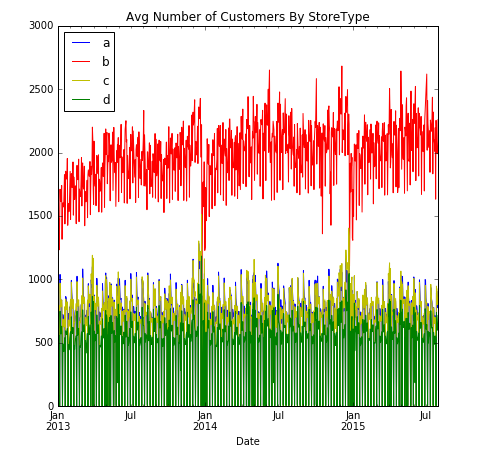


Figure13:

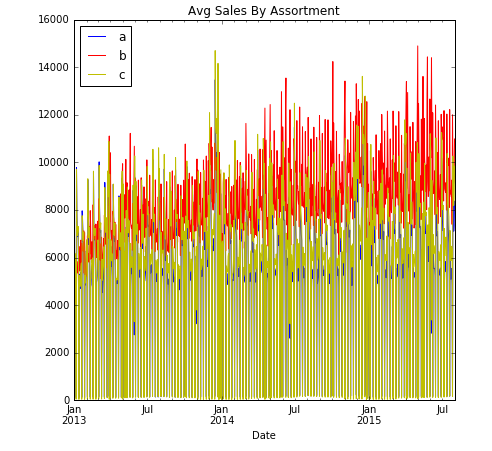


Figure14:

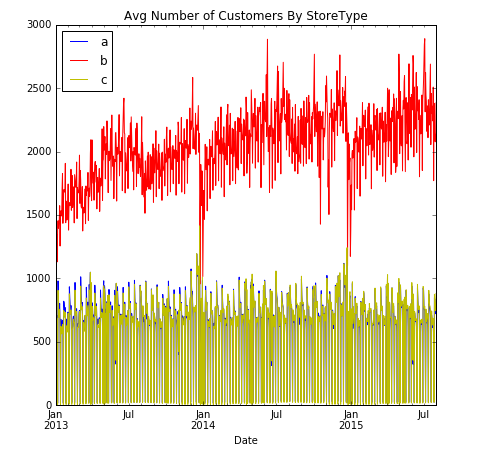


Figure 15:

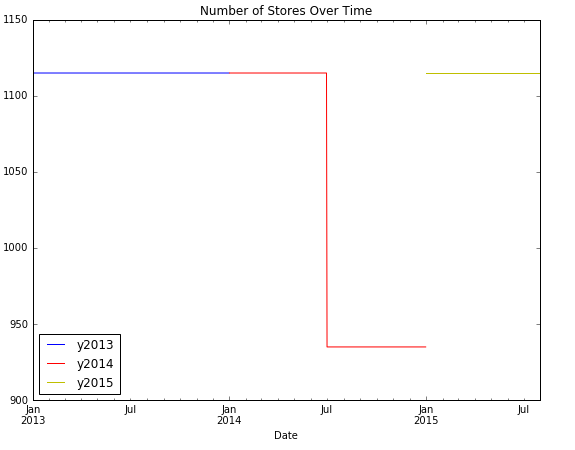
****

Figure 16:

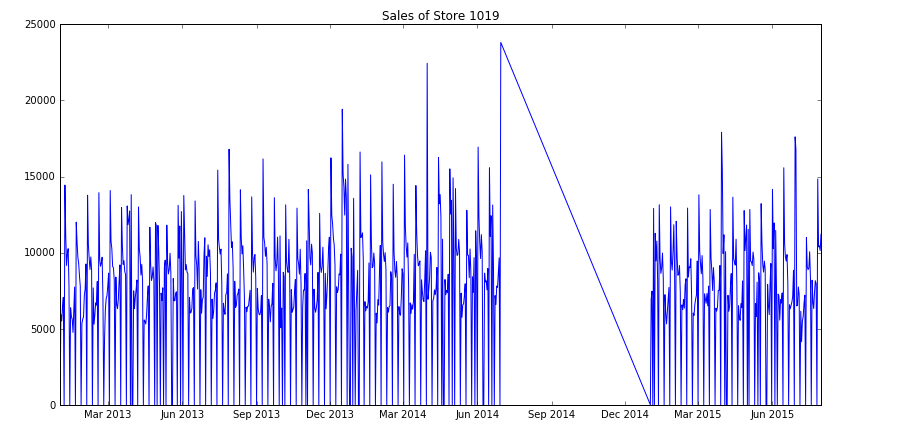


Figure 17:

