Walmart Data Challenge

Shridivya Sharma
shrishar@umail.iu.edu
Masters in Computer Science

October 9, 2017

1 Walmart Data Challenge

Walmart operates 11,450 stores in 27 countries, managing inventory across varying cultures and demographics. In this hack, Walmart challenges participants to assess the impact of promotions and influence of competitors on sales and customer count within stores. Intuitively, we may expect better sales/customer count on days promotions were laid out but the effect is confounded by promotions extended by competitors as well.

2 Data description

We are given two files - **Sales customer** data file contains sales and customers for each store from dates ranging from Jan-2013 to July 2015. It also gives us insights about stores which are opened on State holidays,School holidays and which stores are running promotions. **Store** data gives us information on each store competitors,store assortment type and what type of store it is.

Sales Customer Data:

1. Store - a unique Id for each store
2. Sales - the turnover for any given day (this is what you are predicting)
3. Customers - the number of customers on a given day
4. Open - an indicator for whether the store was open: 0 = closed, 1 = open
5. StateHoliday - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays.
6. SchoolHoliday - indicates if the (Store, Date) was affected by the closure of public schools
7. Promo - indicates whether a store is running a promo on that day

Store Data:

1. Store Type - differentiates between 4 different store models: a, b, c, d
2. Assortment - describes an assortment level: a = basic, b = extra, c = extended
3. CompetitionDistance - distance in meters to the nearest competitor store

4. CompetitionOpenSince[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened

5. Promo2 - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating

6. Promo2Since[Year/Week] - describes the year and calendar week when the store started participating in Promo2

7. PromoInterval - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. ”Feb,May,Aug,Nov” means each round starts in February, May, August, November of any given year for that store

3 Initial Steps:

I decided to use Python 3.3 (Pandas library) to analyse the data set. Before loading the sales customer data, I parsed the date values and divided the data into Training and Test components. Dates from Jan 2013 to May 2015 is considered for training whereas June and July 2015 are considered for testing. I made a copy of the test data as test_actual to calculate the accuracy of our prediction models.

4 Data Visualizations:

Before creating the predictive models, I came up with different visualizations to draw important inferences about our data.

1. Average sales and percent over time Jan 2013 to May 2015. As you can see in Nov and Dec 2013, Dec 2014 and Mar 2015,we had maximum sales. Since its a festive season (thanksgiving and christmas)- **people tend to buy more during that period.**

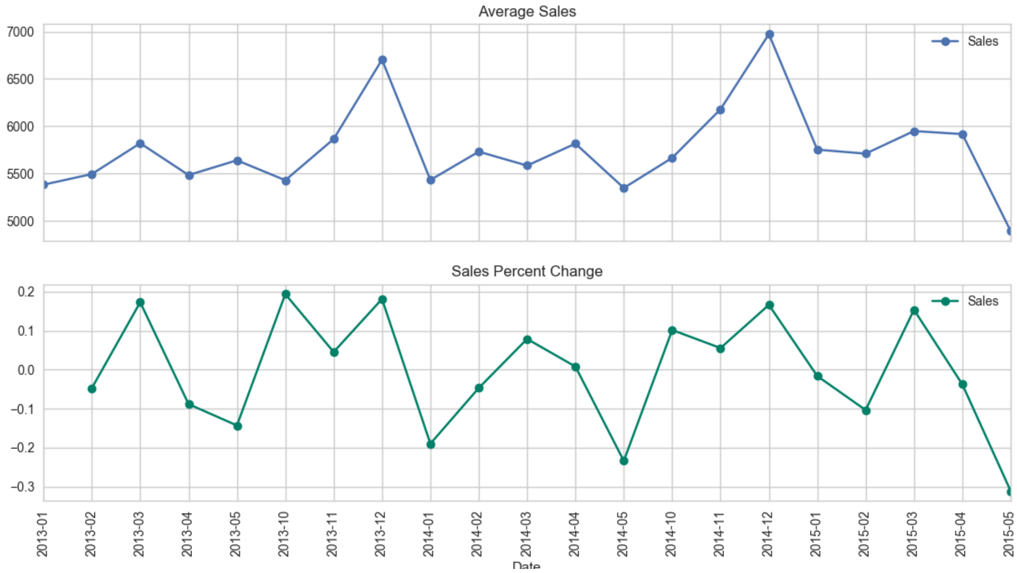


Figure 1: Average Sales over time

2. Average customer over time Jan 2013 to May 2015.As you can see in Nov and Dec 2013, Dec 2014 and Mar 2015,we had maximum customers. It follows the same trend as average sales above.

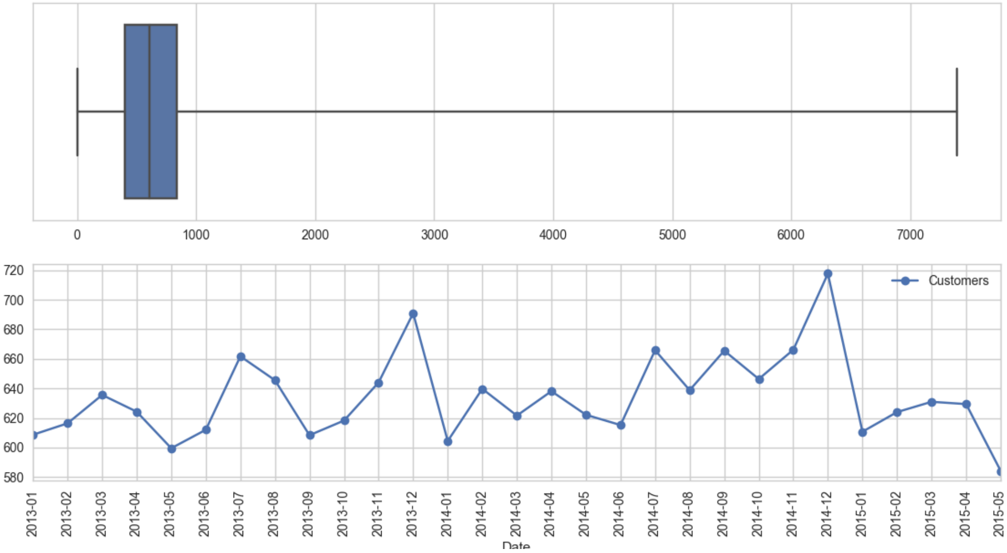


Figure 2: Average Customer over time

3. Customers tend to buy more during the weekdays than weekends. As you can see from the figure below, sales and customer count on sunday is the minimum among all days of week. So it is better to provide promotions on weekdays than weekends.

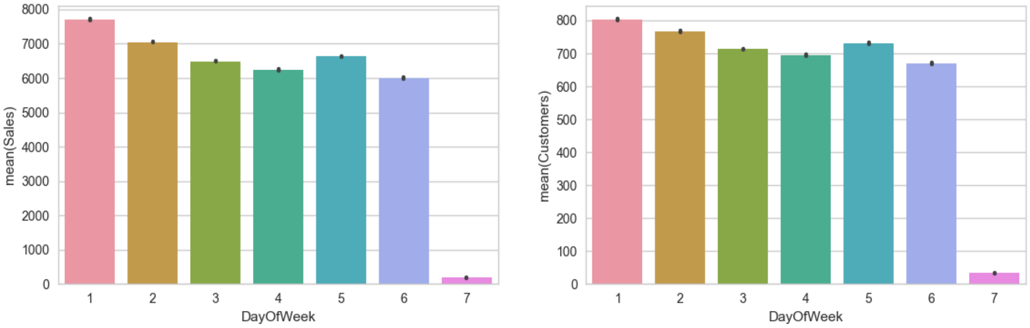


Figure 3: Sales and Customer during week

4. It is clearly visible from the below figure that customer tend to purchase more during promotions. To increase the sales of a store,we should come up with different deals or promotions to attract customers.

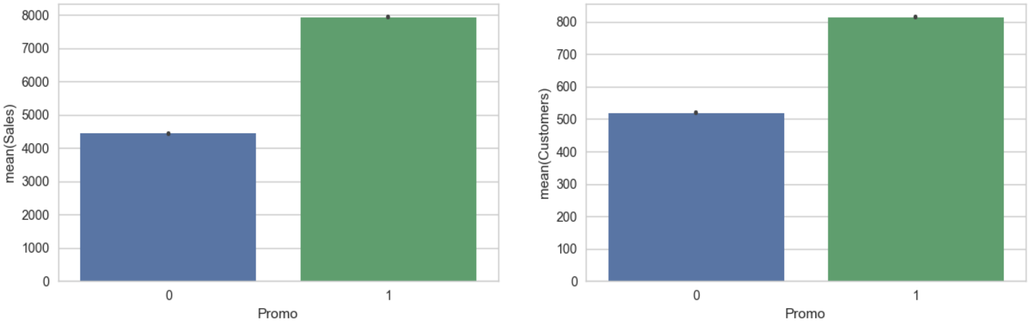


Figure 4: Sales and Customer with/without promo

5. Holidays: We have three state holiday data with us - public holiday, easter holiday and christmas holiday. In general, a state/public holiday usually ends up with a long weekend and customers tend to shop before the start of vacations. Fig 9 shows some statistics about school holidays. School holidays plays an important role in increasing sales since kids will be at home and will perform lot of summer activities during holidays. Parents along with the kids would visit the stores more often than usual. So during this period,parents might end up buying things for themselves apart from buying to the kids. This can be a potential profit.

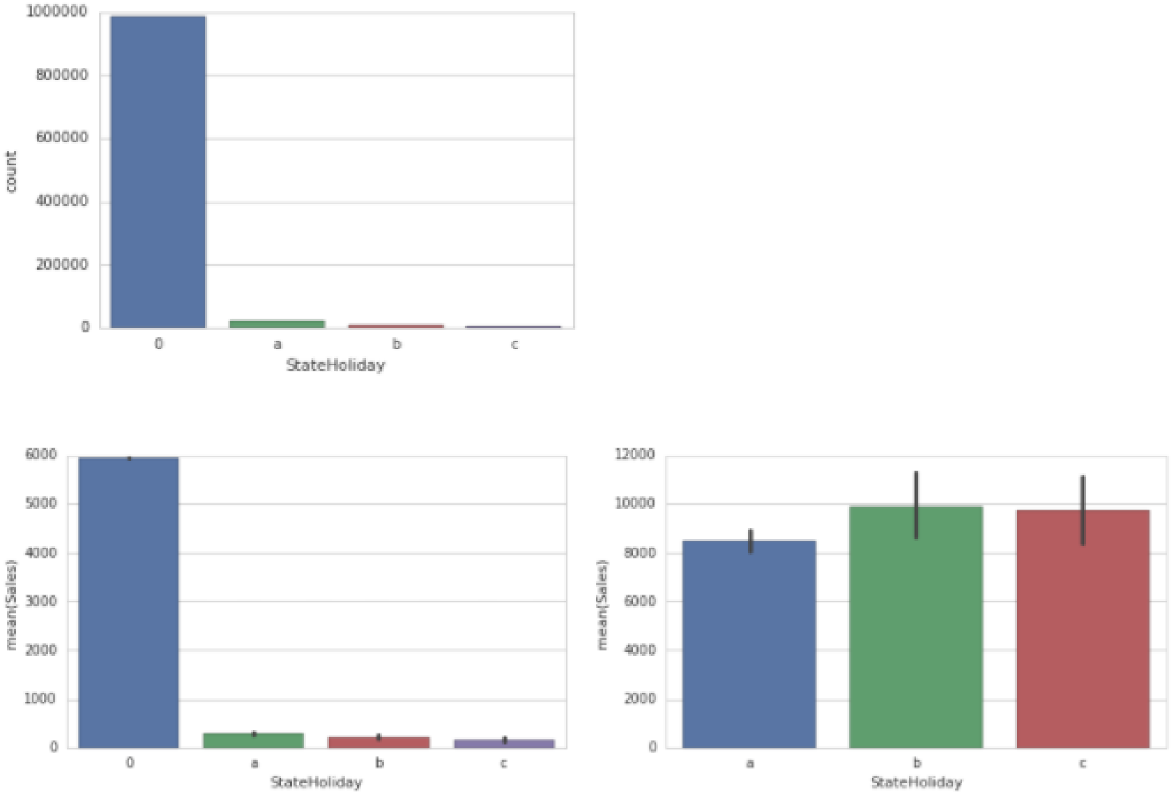


Figure 5: State holidays Sales and Customer

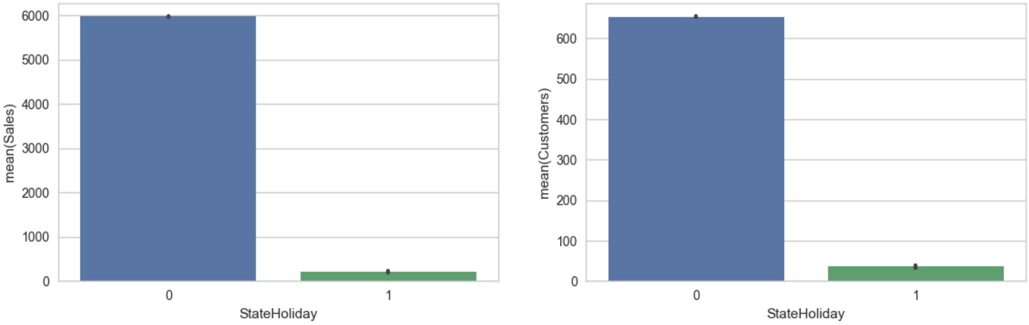


Figure 6: State holidays Sales and Customer when store was opened and closed

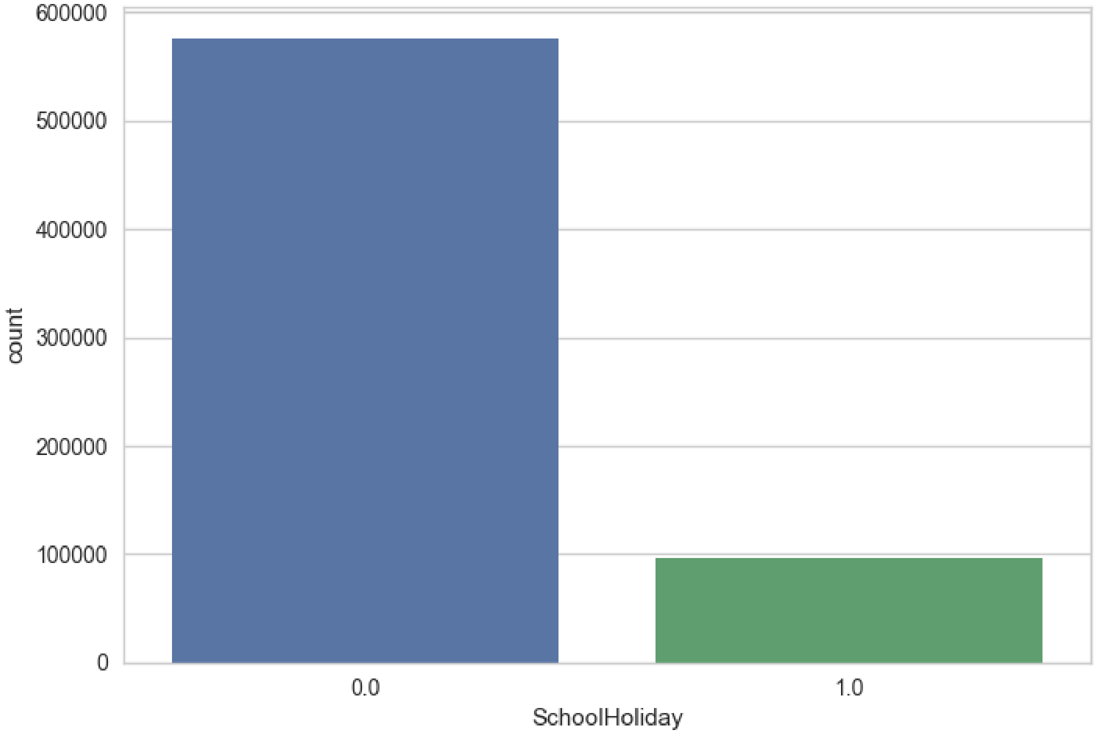


Figure 7: School holidays customer count on open and closed days

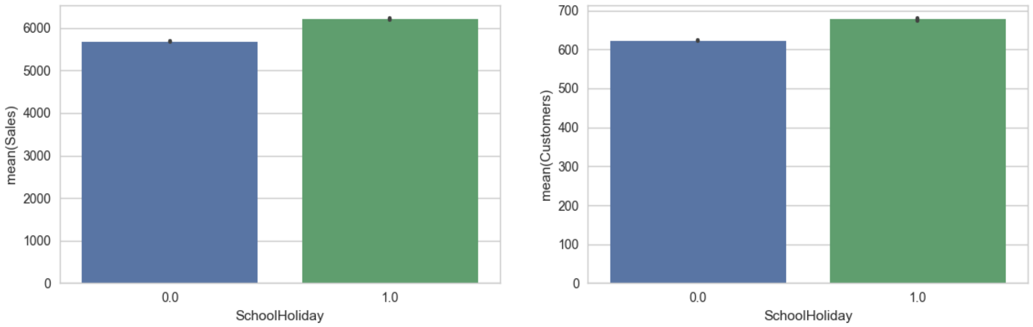


Figure 8: School holidays Sales and Customer

6. Store type and Assortment type : This visualization is basically used to check what kind of stores and assortment type customers like to go for. As you can see below, customers tend to visit store type b and assortment b i.e. basic

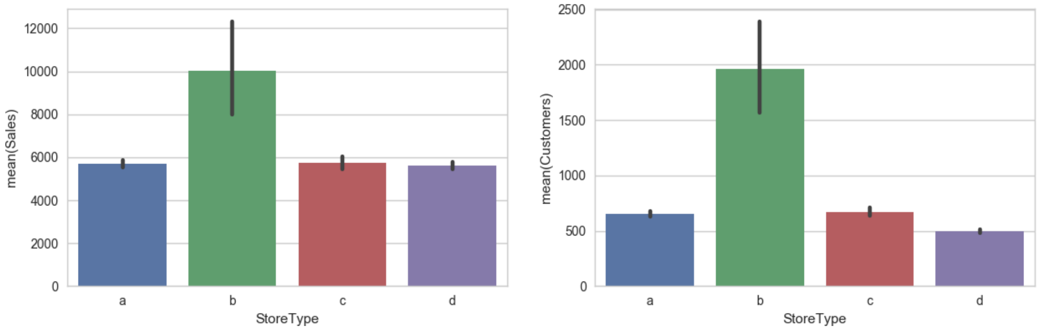


Figure 9: Average sales and customers for each store type

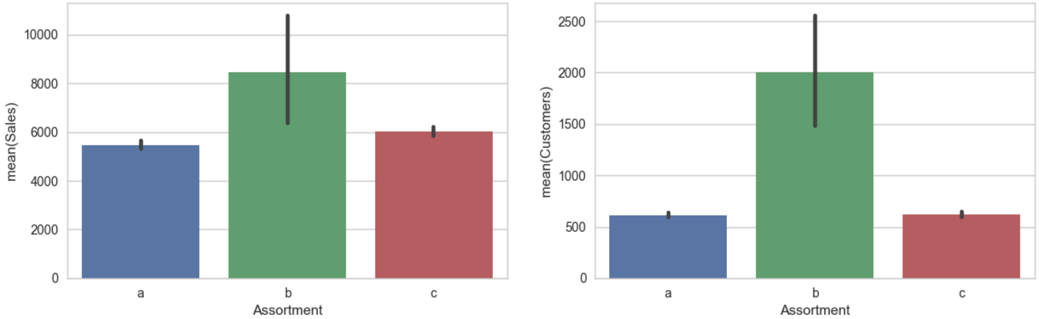


Figure 10: Average sales and customers for each assortment type

7. Competitor: If a store is closeby to a walmart store which gives a better deals or promotions, customers tend to visit the other stores.Let's look at the competitors distance:

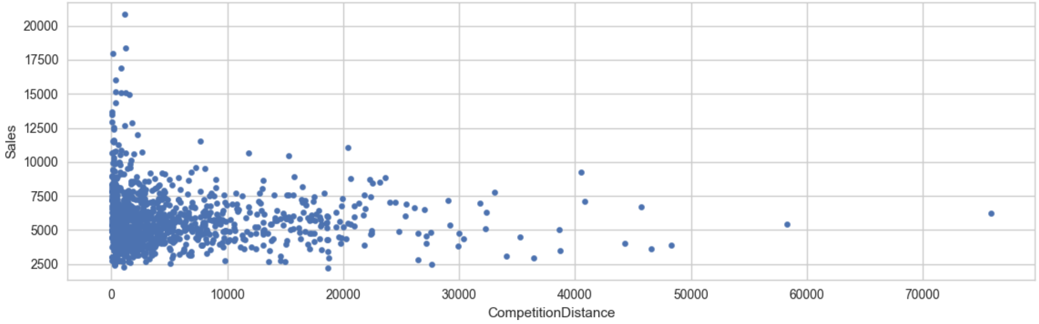


Figure 11: Competitors distance

As you can see , most of the competitors are close by to the walmart stores.

8. Competitor Effect: Let's look at the competitor effect on the Walmart stores. Consider Figure 13 for example. Store 2 did a decent job in sales during Christmas and thanksgiving but relatively the sales decreased after the festival season. The competitor for Store 2 is within 570 meters and they are offering promotions as well (as seen from the data). On the contrary, Store 1 has a competitor within 1270 meters and if you look at the sales of Store 1, they peak during festival season(as expected) but remain constant during the rest of the year.

   Store 2 doesn't do well as compared to Store 1, and one of the main reasons for it can be the distance of the competitor.

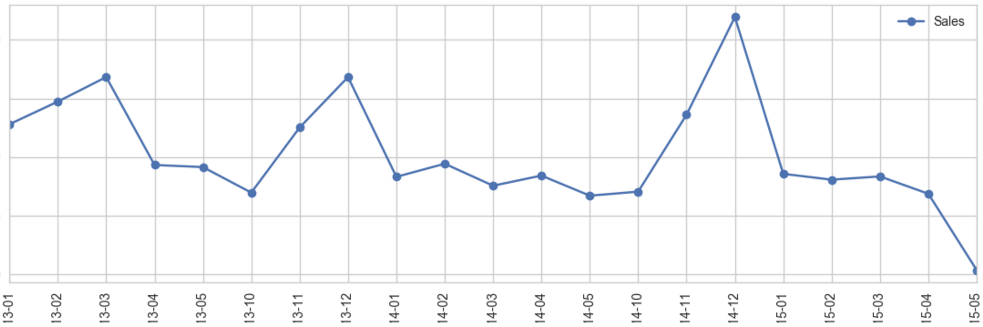


Figure 12: Competitors Effect on Store 1

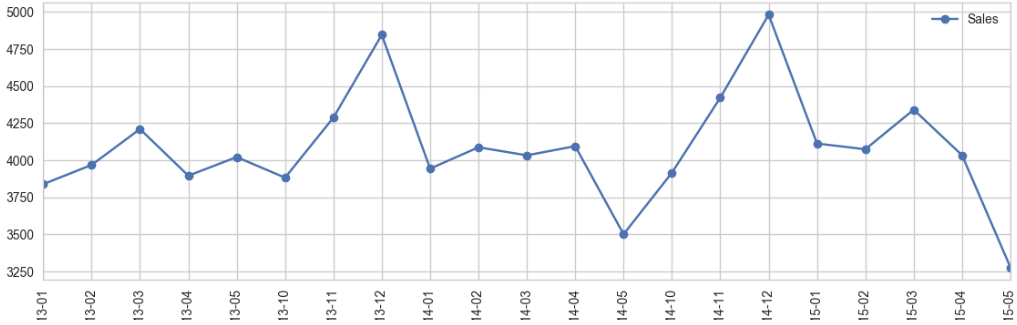


Figure 13: Competitors Effect on Store 2

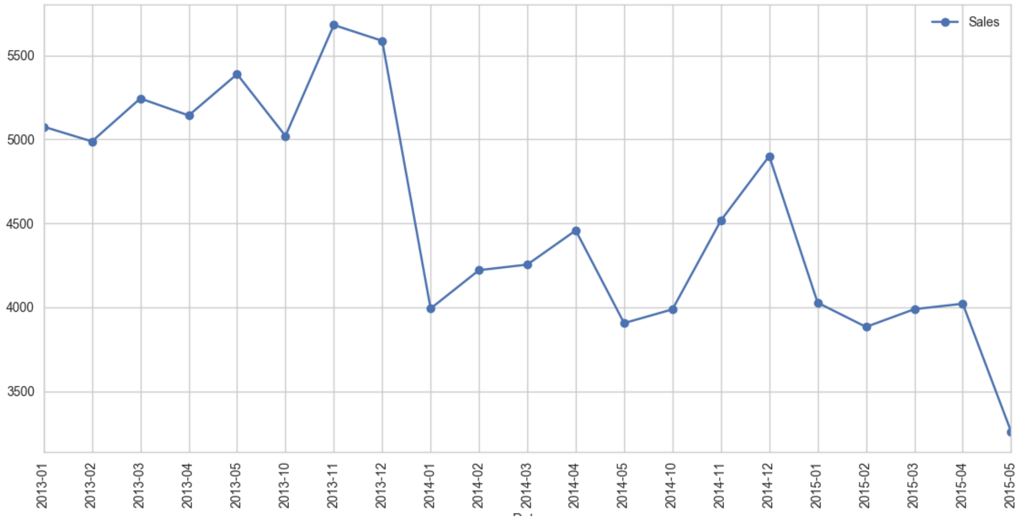


Figure 14: Competitors Effect on Store 6

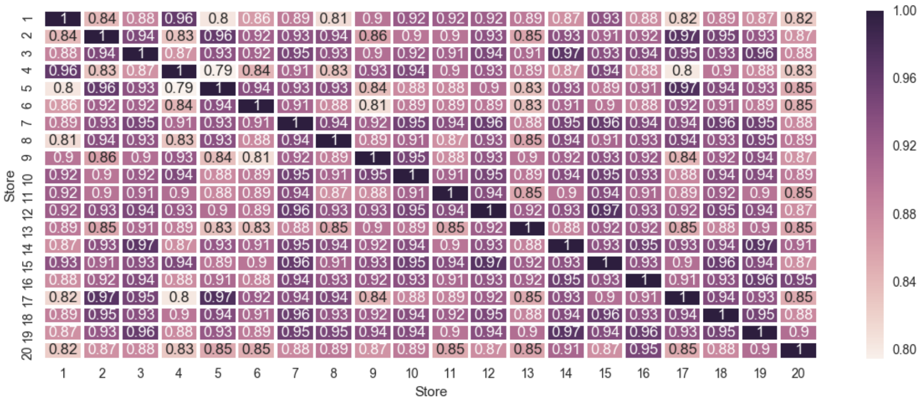
5 Data Preprocessing:

I combined training and store data with a join on Store id (store is common between salescust and store dataframe). Steps:

1. We are interested in sales of stores which were open. So I discarded the values when stores were closed.
2. Date field is further divided into three columns : Year, Month and week of year. Later the actual date field was dropped.
3. I checked the competitors for each store. If there is a competitor for a store,I replaced the value as 1 else 0.
4. I then checked for promo2 value for each store and followed the same technique as above.
5. All the categorical values for assortment type,store type and promo interval is hot encoded.
6. I dropped Promo2SinceYear, Promo2SinceWeek,CompetitionOpenSinceMonth, CompetitionOpenSinceYear since it didn't look necessary after what I saw from the visualizations.

5.1 Collinearity:

Before modeling the data,I checked collinearity between the range of stores in our data set. Results are as follows: There is a lot of collinearity between the attributes in the data set.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	0.84	0.88	0.96	0.8	0.86	0.89	0.81	0.9	0.92	0.92	0.92	0.89	0.87	0.93	0.88	0.82	0.89	0.87	0.82
2	0.84	1	0.94	0.83	0.96	0.92	0.93	0.94	0.86	0.9	0.9	0.93	0.85	0.93	0.91	0.92	0.97	0.95	0.93	0.87
3	0.88	0.94	1	0.87	0.93	0.92	0.95	0.93	0.9	0.92	0.91	0.94	0.91	0.97	0.93	0.94	0.95	0.93	0.96	0.88
4	0.96	0.83	0.87	1	0.79	0.84	0.91	0.83	0.93	0.94	0.9	0.93	0.89	0.87	0.94	0.88	0.8	0.9	0.88	0.83
5	0.8	0.96	0.93	0.79	1	0.94	0.93	0.93	0.84	0.88	0.88	0.9	0.83	0.93	0.89	0.91	0.97	0.94	0.93	0.85
6	0.86	0.92	0.92	0.84	0.94	1	0.91	0.88	0.81	0.89	0.89	0.89	0.83	0.91	0.9	0.88	0.92	0.91	0.89	0.85
7	0.89	0.93	0.95	0.91	0.93	0.91	1	0.94	0.92	0.95	0.94	0.96	0.88	0.95	0.96	0.94	0.94	0.96	0.95	0.88
8	0.81	0.94	0.93	0.83	0.93	0.88	0.94	1	0.89	0.91	0.87	0.93	0.85	0.94	0.91	0.93	0.94	0.93	0.95	0.89
9	0.9	0.86	0.9	0.93	0.84	0.81	0.92	0.89	1	0.95	0.88	0.93	0.9	0.92	0.93	0.92	0.84	0.92	0.94	0.87
10	0.92	0.9	0.92	0.94	0.88	0.89	0.95	0.91	0.95	1	0.91	0.95	0.89	0.94	0.95	0.93	0.88	0.94	0.94	0.89
11	0.92	0.9	0.91	0.9	0.88	0.89	0.94	0.87	0.88	0.91	1	0.94	0.85	0.9	0.94	0.91	0.89	0.92	0.9	0.85
12	0.92	0.93	0.94	0.93	0.9	0.89	0.96	0.93	0.93	0.95	0.94	1	0.92	0.93	0.97	0.93	0.92	0.95	0.94	0.87
13	0.89	0.85	0.91	0.89	0.83	0.83	0.88	0.85	0.9	0.89	0.85	0.92	1	0.88	0.92	0.92	0.85	0.88	0.9	0.85
14	0.87	0.93	0.97	0.87	0.93	0.91	0.95	0.94	0.92	0.94	0.9	0.93	0.88	1	0.93	0.95	0.93	0.94	0.97	0.91
15	0.93	0.91	0.93	0.94	0.89	0.9	0.96	0.91	0.93	0.95	0.94	0.97	0.92	0.93	1	0.93	0.9	0.96	0.94	0.87
16	0.88	0.92	0.94	0.88	0.91	0.88	0.94	0.93	0.92	0.93	0.91	0.93	0.92	0.95	0.93	1	0.91	0.93	0.96	0.95
17	0.82	0.97	0.95	0.8	0.97	0.92	0.94	0.94	0.84	0.88	0.89	0.92	0.85	0.93	0.9	0.91	1	0.94	0.93	0.85
18	0.89	0.95	0.93	0.9	0.94	0.91	0.96	0.93	0.92	0.94	0.92	0.95	0.88	0.94	0.96	0.93	0.94	1	0.95	0.88
19	0.87	0.93	0.96	0.88	0.93	0.89	0.95	0.95	0.94	0.94	0.9	0.94	0.9	0.97	0.94	0.96	0.93	0.95	1	0.9
20	0.82	0.87	0.88	0.83	0.85	0.85	0.88	0.89	0.87	0.89	0.85	0.87	0.85	0.91	0.87	0.95	0.85	0.88	0.9	1

Figure 15: Pearson coefficient

6 Data Analysis:

After preprocessing the data,I came up with different models for predicting Sales value for the month of June and July 2015 (test data). Here are the few data models that I used to predict the values:

Data model	r^2 score
RandomForestRegressor	1.428
RidgeRegressor	2.8648
Lasso Regression	3.455

7 Conclusion

Below are the major observations -

1. It is stated that major sales happen during promotions and we can verify the same with our data as shown in Figure 5.
2. Maximum sales occur in December and end of November during Christmas and Thanksgiving, so promotions play a vital role during these holidays.
3. Maximum sales occur during weekdays, hence promotions can be awarded on weekdays rather than weekends.
4. Whenever state holidays end up as long weekends, sales increases prior to the start of holidays. So it is advisable to have promotions prior to the start of long weekends or state holidays.
5. Customers tend to visit store type 'b' and tend to purchase assortment 'b' more. More promotions at these stores and on these assortments will result in better sales.
6. General trend about Competitors is greater the distance to the competitor, better the sales.

This is the best I could come up with, given the limited timeline and other factors(had a midterm at my school). Given more freedom, I would have tried the below methods to further improve the accuracy of the predictive model and to better analyse the data -

1. Use Auto-regressive Integrated Moving Average (ARIMA) to generate time series models, which will help to select specific set of features for the regression models. Seasonal ARIMA could provide us a
2. Use different feature reduction techniques to analyse what features play critical role while constructing the model. I am currently using all the features provided in the data set.
3. Play around with different regression algorithms or statistical methods to better fit the data and in turn yield better accuracy.
4. Much better visualizations, using Tableau or Plotly.