WALMART SALES FORECASTING

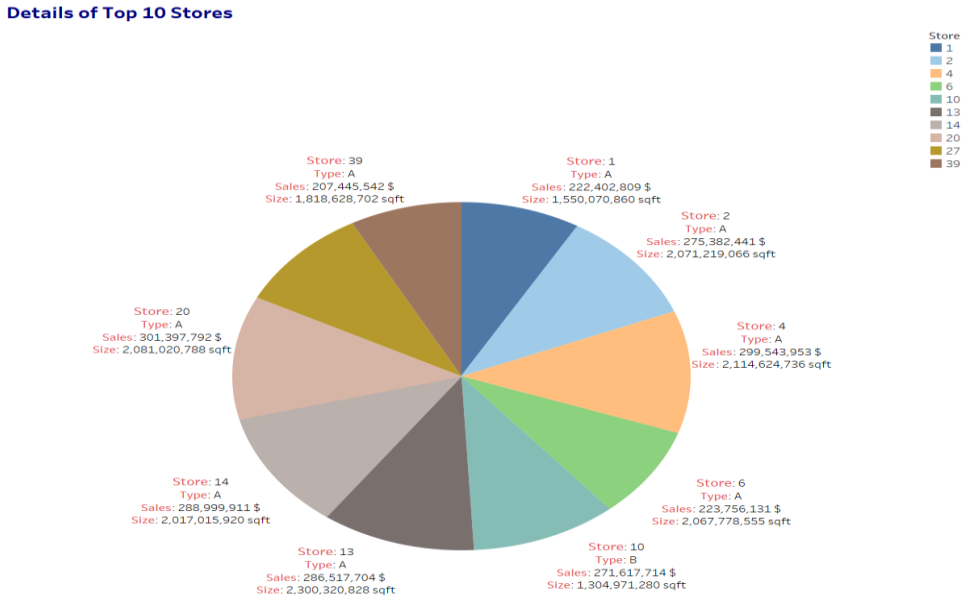
Source for dataset and problem statement:

https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data

Introduction

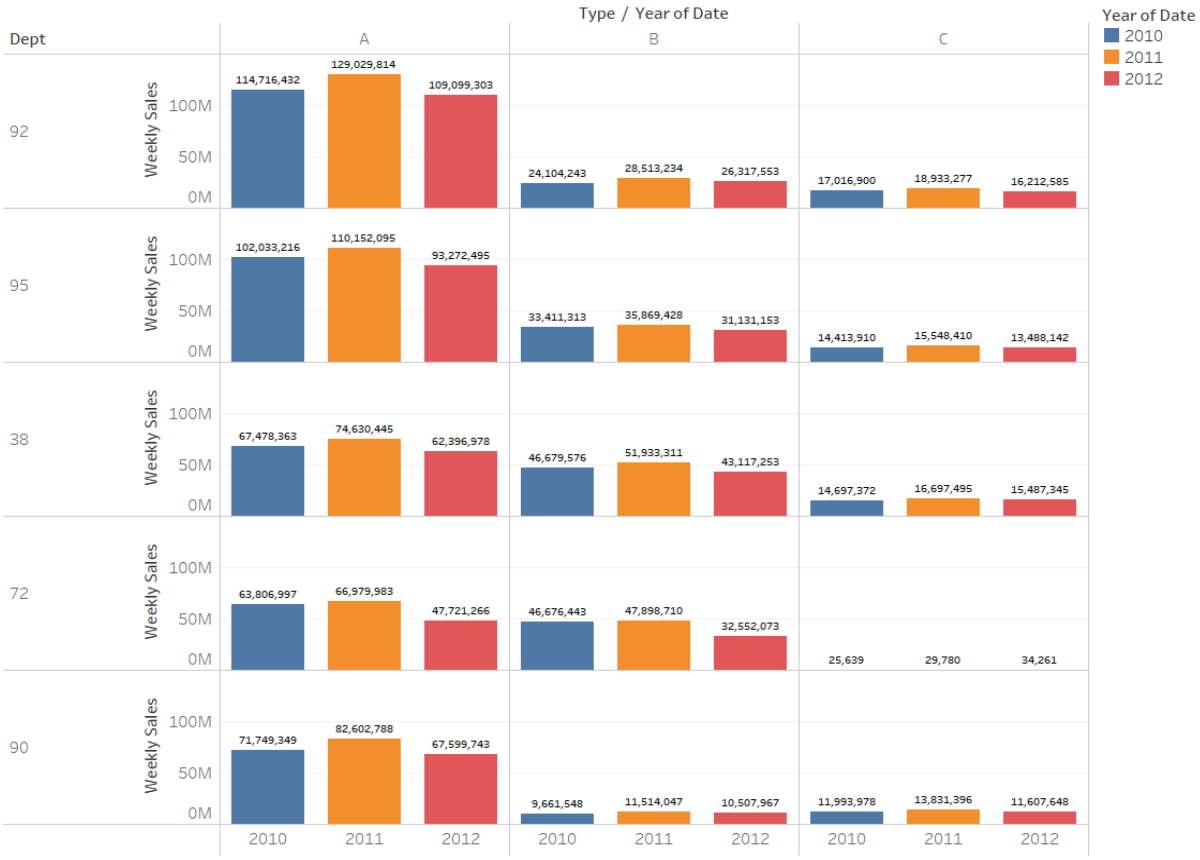
- Super store, Walmart has a huge impact on their sales during markdown events like the Super Bowl,President's day and other holidays like Black Friday, Christmas,etc.
- The reason being there are several promotional offers preceding these holidays.
- Therefore, there is a necessity to analyze the sales data set of the MegaStore Walmart, during markdown events in order to plan more efficient strategies for improving sales

Exploratory Analysis



Store (stores.csv), Type, sum of Weekly Sales and sum of Size. Color shows details about Store. The marks are labeled by Store (stores.csv), Type, sum of Weekly Sales and sum of Size. The view is filtered on Store, which has multiple members selected.

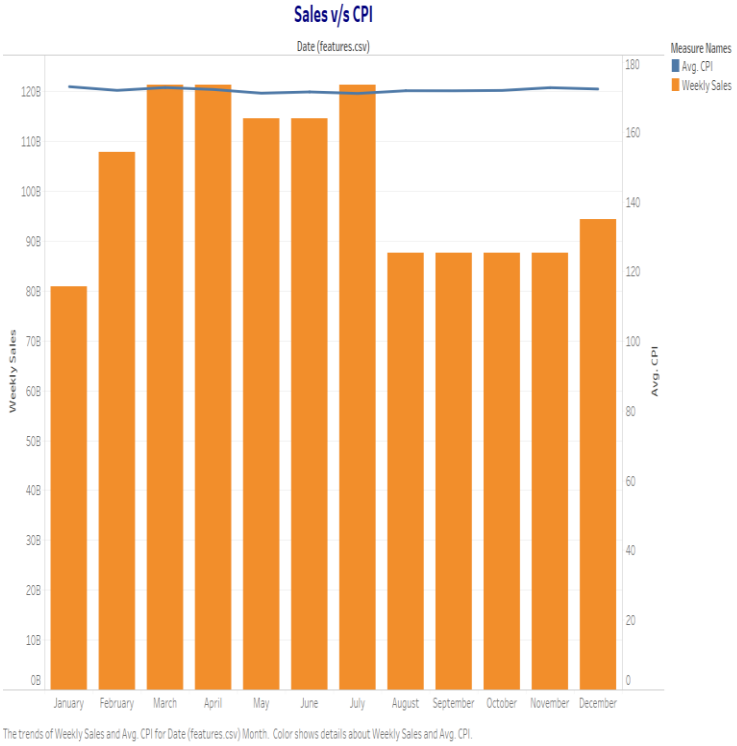
TOP 5 Departments across the Stores



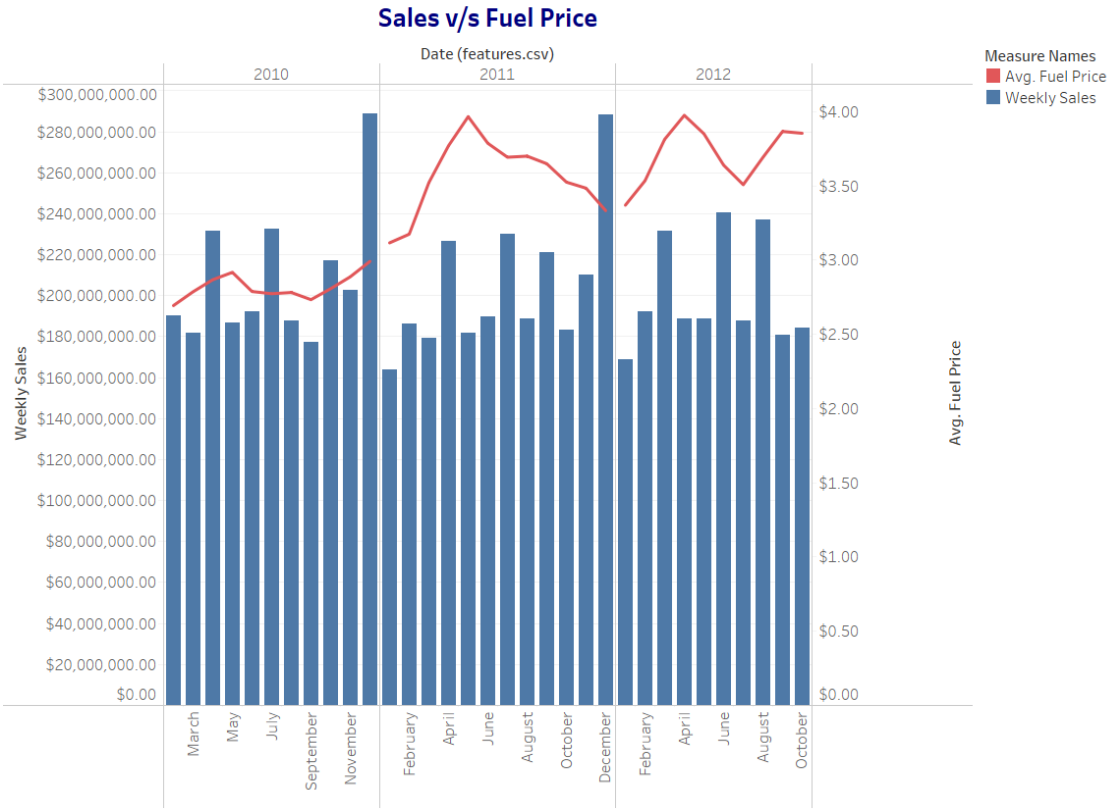
Sum of Weekly Sales for each Date Year broken down by Type vs. Dept. Color shows details about Date Year. The marks are labeled by sum of Weekly Sales. The view is filtered on Dept, which keeps 38, 72, 90, 92 and 95.

Store Number	Cumulative Weekly Sales($)
92	483,943,342
95	449,320,163
38	393,118,137
72	305,725,152
90	291,068,469

- No strong relationships were clear from visualizing the weekly sales data with respect to the CPI and the fuel price during that week.



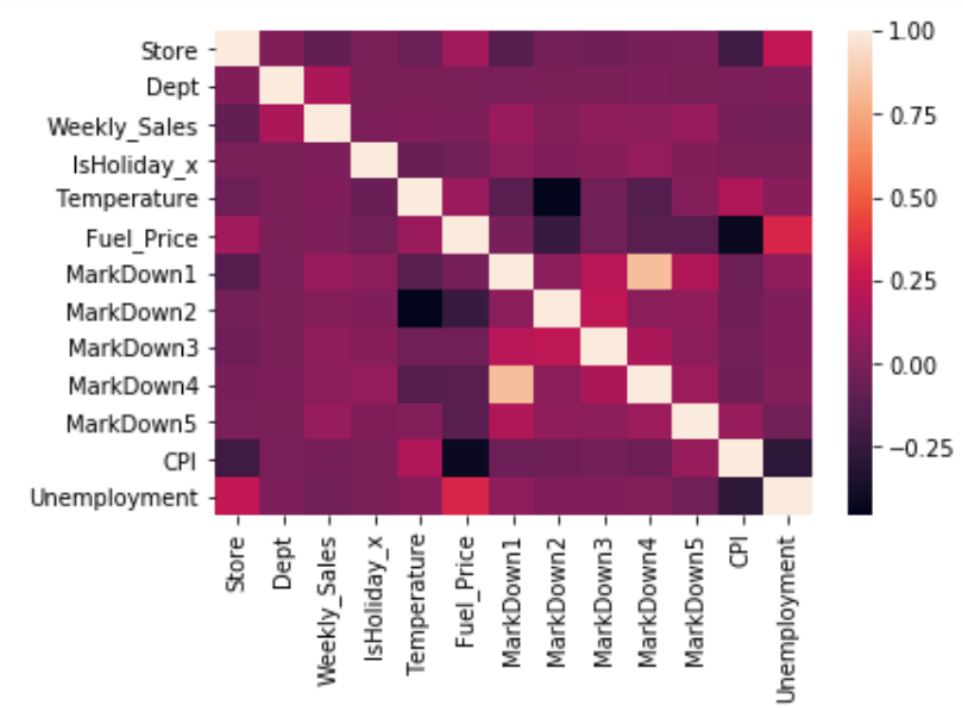
The trends of Weekly Sales and Avg. CPI for Date (features.csv) Month. Color shows details about Weekly Sales and Avg. CPI.



The trends of Weekly Sales and Avg. Fuel Price for Date (features.csv) Month broken down by Date (features.csv) Year. Color shows details about Weekly Sales and Avg. Fuel Price.

Autocorrelation Heat Map

	Store	Dept	Weekly_Sales	IsHoliday_x	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment
Store	1	0.02	-0.087	9e-05	-0.053	0.13	-0.13	-0.028	-0.042	-0.01	-0.0046	-0.21	0.24
Dept	0.02	1	0.16	0.00066	0.0028	0.00059	-0.0017	-0.00073	0.00066	0.0049	0.0012	-0.0066	0.0052
Weekly_Sales	-0.087	0.16	1	0.0079	0.015	0.0072	0.1	0.031	0.071	0.06	0.092	-0.018	-0.036
IsHoliday_x	9e-05	0.00066	0.0079	1	-0.059	-0.032	0.061	0.013	0.041	0.087	0.02	-0.0008	0.003
Temperature	-0.053	0.0028	0.015	-0.059	1	0.11	-0.12	-0.46	-0.033	-0.13	0.033	0.18	0.044
Fuel_Price	0.13	0.00059	0.0072	-0.032	0.11	1	-0.016	-0.24	-0.034	-0.13	-0.12	-0.42	0.33
MarkDown1	-0.13	-0.0017	0.1	0.061	-0.12	-0.016	1	0.047	0.2	0.83	0.18	-0.056	0.068
MarkDown2	-0.028	-0.00073	0.031	0.013	-0.46	-0.24	0.047	1	0.23	0.048	0.07	-0.049	0.013
MarkDown3	-0.042	0.00066	0.071	0.041	-0.033	-0.034	0.2	0.23	1	0.15	0.056	-0.029	0.015
MarkDown4	-0.01	0.0049	0.06	0.087	-0.13	-0.13	0.83	0.048	0.15	1	0.11	-0.05	0.029
MarkDown5	-0.0046	0.0012	0.092	0.02	0.033	-0.12	0.18	0.07	0.056	0.11	1	0.095	-0.03
CPI	-0.21	-0.0066	-0.018	-0.0008	0.18	-0.42	-0.056	-0.049	-0.029	-0.05	0.095	1	-0.28
Unemployment	0.24	0.0052	-0.036	0.003	0.044	0.33	0.068	0.013	0.015	0.029	-0.03	-0.28	1



Process:

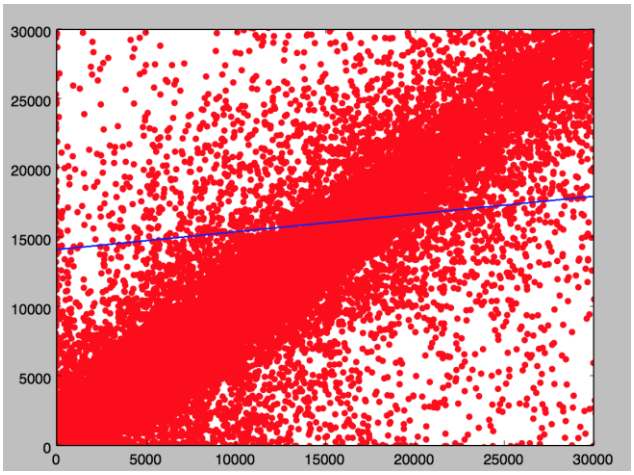
Data Preparation: Merging, Cleaning, and Transforming the Data

- We are merging train.csv and features.csv based on Store numbers and Date.
- From the Merged Data, we are removing records dated from Jan 2011 to November 2011, since the markdowns(1-5) in features.csv for this period is unavailable.
- For the remainder of Merged Data, we found missing values in Markdown(1-5) columns. We are filling these missing values with mean grouped by store and department.
- Test.csv consists of Store, Department and Date columns dated from November 2012 to July 2013.
- Only for the computation of accuracy score, we are dividing the train data in 70:30 ratio.

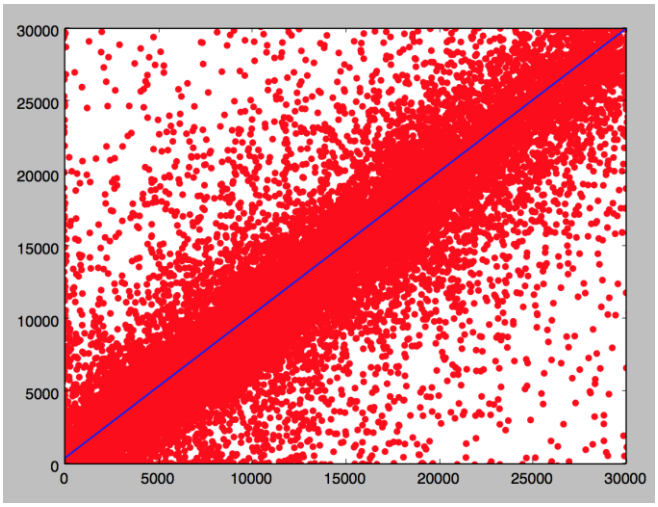
Supervised Modelling

	Linear Regression	Random Forest Regressors	Extra Trees Regressors.
R2	-5.43	0.86	0.87
Mean Log Square error	0.91	0.67	0.65

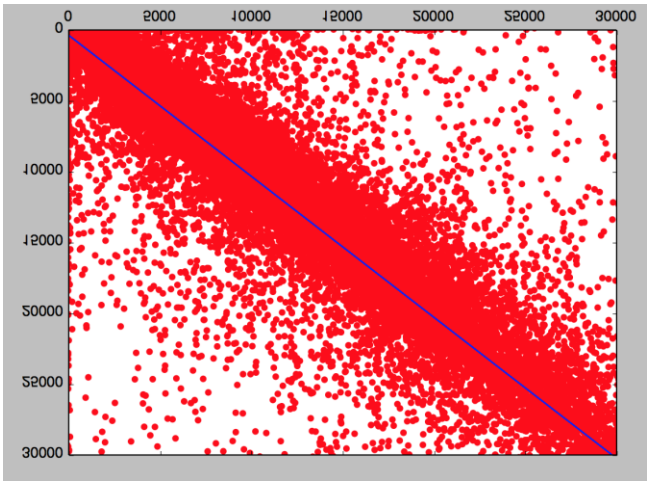
- **Linear Regression**



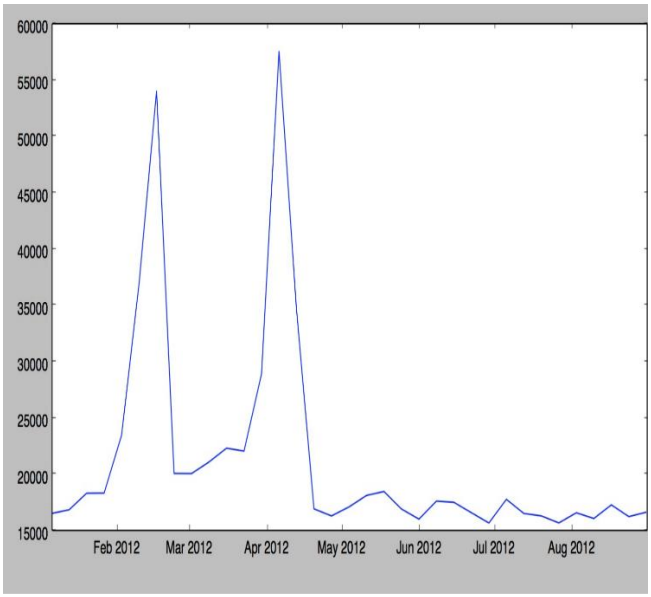
Random Forest Regressors

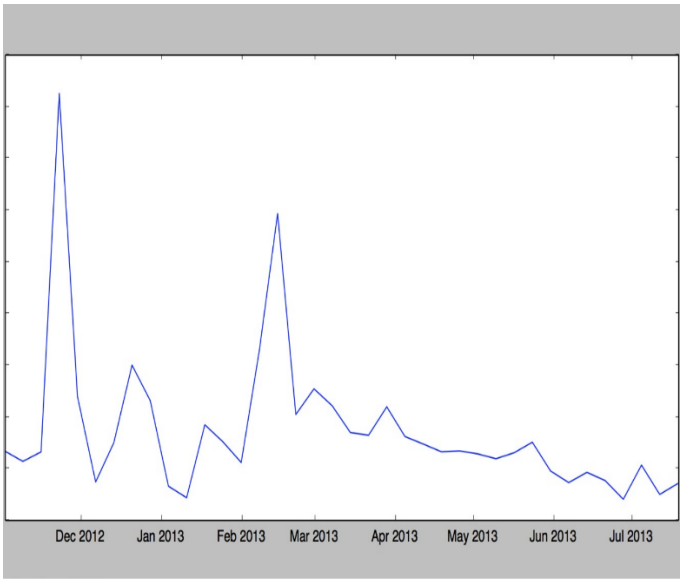


Extra Trees Regressors



Sales Trend Analysis on Modeled Data





Conclusion and Improvements

- Performance : Random Forest Regressors, Random Forest Regressors > Linear Regression
- Sales :Sales increase during the months February, April, November and December which is in accordance with our Trend Analysis.