Project: Predicting sales

*You may work in teams of two. Hand in by end of business day, 21 March, at Gleacher Center Suite 430 or Harper Center Suite 448*

Recently (20 February 2014 to 5 May 2014), Walmart recruited analysts via an online competition at kaggle.com:


Here is the description of this competition.

One challenge of modeling retail data is the need to make decisions based on limited history. If Christmas comes but once a year, so does the chance to see how strategic decisions impacted the bottom line.

In this recruiting competition, job-seekers are provided with historical sales data for 45 Walmart stores located in different regions. Each store contains many departments, and participants must project the sales for each department in each store. To add to the challenge, selected holiday markdown events [“Clearance”, “Rollbacks”, “Special Buys”] are included in the dataset. These markdowns are known to affect sales, but it is challenging to predict which departments are affected and the extent of the impact.

For our final project this year, we will be analyzing this same data set. The data is available at the class web page:


For the official contest, sales data was withheld from participants and Walmart used this data to evaluate the submitted prediction models. Here, we will undertake a more informal analysis, simply trying to explore the provided data to unearth insights about the relationship between weekly sales and various factors.
Specifically, the provided factors are:

- **Store_ID**
- **Date**: calendar week
- **Dept**: the department number
- **Temperature**: average temperature in the region
- **Fuel_Price**: cost of fuel in the region
- **Markdown1** thru **Markdown5**: data related to promotional markdowns
- **CPI**: consumer price index
- **Unemployment**: the unemployment rate
- **IsHoliday**: whether the week is a special holiday week

This analysis is intentionally open ended. While you explore the data, recall the tools you have learned in class, including (but not limited to):

- frequency tables (sometimes called pivot tables)
- conditional probabilities and conditional expectations
- scatter plots
- density plots and histograms
- linear regression and prediction intervals
- hypothesis tests and confidence intervals
- simulation based permutation tests

The maximum length of your report should be fifteen pages, including plots and tables. The general format should be an *introduction* outlining the aspects you plan to explore and describing the data, the *analysis* itself, including plots, diagrams, and tables, and a *discussion/conclusion* section summarizing your findings.

In class of week ten we will work on this project, so be prepared with questions.

Good luck!