# Midterm Report

Vidhan Bhatt (vpb24)
Steven Jaroslawski (sj393)

October 28, 2016

## Project Background

We would like to forecast Walmart's future sales by store and department for 45 stores. This is a valuable problem to solve from an inventory management, pricing, and staffing perspective. Walmart executives would like to know how much inventory to carry by department from week to week, well in advance. Store managers can utilize the information to determine employee shift schedules that are optimized for weekly demand. And lastly, the pricing department will gain insight into which markdowns are effective and which are not.

This is an especially important task because Walmart generates a bulk of its revenue from strategic markdowns (discounts), the four largest of which precede major events/holidays in the United States (Super Bowl, Labor Day, Thanksgiving, Christmas). This phenomenon poses a challenge: modeling the impact of these markdowns with sparse data because each event only occurs annually.

## Description of Data Sets

We are using data from Kaggle. There are four notable files that we will use: stores.csv (stores) features.csv (features), train.csv (training data), and test.csv (test data).

The stores data contains one row for each of the 45 stores, and two columns for each store: the type of store, classified into three buckets (which are presumably Walmart Supercenters, Walmart Discount Stores, and Walmart Neighborhood Markets, the company's three most popular store types), and the square footage of each store.

The features data contains macroeconomic indicators for the region in which each store is located, for each week from February 5, 2010 to November 01, 2012. These indicators are the average temperature, average fuel price, average consumer price index, and average unemployment rate. Additionally, this file contains markdown data for five promotions that Walmart ran. This data is anonymized, only available after November 2011, and only in select stores each week. Besides this, we are only missing average temperatures and CPI values for several weeks in 2013, which will not be included in the training set. The features data also has a boolean column indicating whether each week is a Walmart-designated holiday week or not.

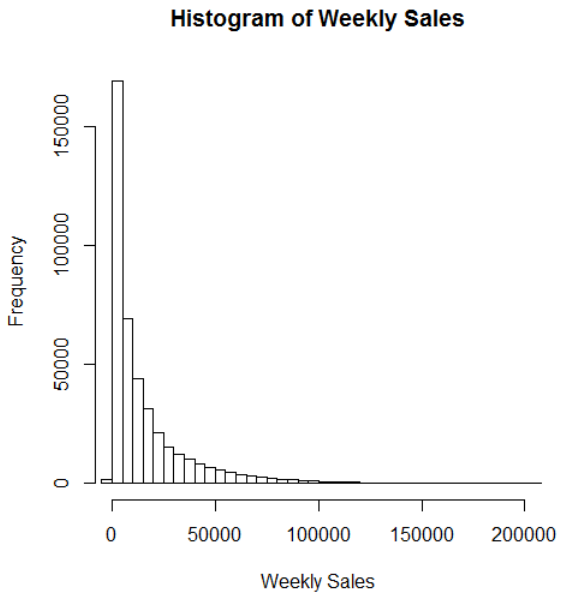
The training data contains weekly sales data for each department within each store, and a boolean column indicating whether each week is a Walmart-designated holiday week not.

We plan to run multiple hypotheses only on the training set to determine which model performs the best. We will be sure to choose enough models so that the best model is not under fit to the data.

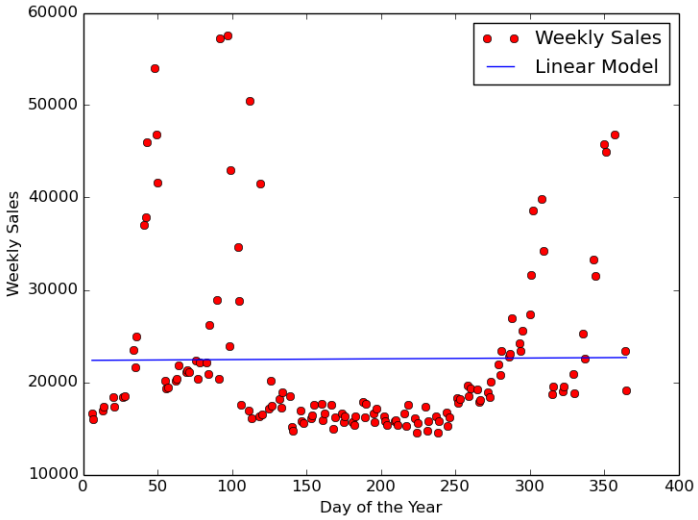
Lastly, the test data contains all the same information as the training data except for weekly sales (which we will predict) for every week from November 2, 2012 to July 26, 2013. We will

run the model that performed the best on the training set on the test set in order to determine its effectiveness. We only run this model to avoid over fitting.

## Preliminary Data Analysis

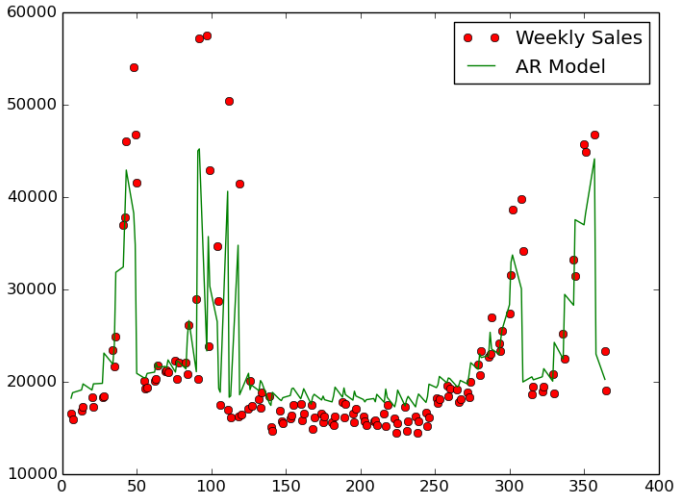



This histogram shows the general shape of our data. The data appears to be exponentially distributed in general, with about 40 percent of the weekly sales less than 5000 dollars and a maximum value of 693099.4. It should be noted that there are 1285 negative values, the most negative value -4988.94, which are most likely bad data but there is a chance these values are due to returns, refunds or something similar.




We first attempted to fit a very simple linear model which only regressed on the day of the

year. This graph shows the model's performance on one department in one store. As expected, this model did not predict very well.




We next chose to use an auto-regressive model. Again, this graph shows the model's output on one department in one store. We thought that an accurate prediction of future weekly sales would be past values of that variable. Further, autoregressive models are especially good at handling flexibility and a wide range of time series patterns. Our data is relatively noisy with high variance during holidays, so the autoregressive model was an appropriate choice. It appeared to predicted the data fairly well.

## What's Left

We have begun creating a few models that we believe are good predictors of sales from the data we have. We will continue refining and testing until we are confident we have exhausted enough possibilities to move onto the test set. We are considering incorporating another data set that provides some more economic and demographic factors that influence Walmart's customers' purchase behavior. Perhaps average income by region and availability of alternative shopping centers, among others, will be valuable predictors that we can integrate.