# Improving Demand Forecasting

Milestone Report, Capstone Project 2

## 1. Problem Statement

What is the seasonality of products in a store? How does demand compare across different stores for the same item? What about seasonally? What will the demand for a product be in the next few months? Accurate forecasting of sales and demand is an important part to managing the supply chain for both online and physical retail. This study in data science analyzes the sales of 50 different items at 10 stores across 5 years and goes through a process of fine tuning the forecasts in demand.

## 2. Exploratory Data Analysis

The dataset comes from Kaggle's store item demand forecasting challenge. It describes the sales numbers of 50 items across 10 stores and comes in 2 csv files, train.csv and test.csv.

### 2.1 Overall Sales

The training dataset contains the date, store number, item number, and sales number for each day of 2013 through 2017, totaling 913,000 records. The dataset contains no details about the items or the stores. This analysis assumes that the stores are from the US.

Daily sales numbers range from 0 to 231 with a mean of 52. The majority, or middle 50%, of daily item sales numbers fall between 30 and 70, and there are the least number of observations above 100. Density of the sales observations shows the data has non-normal distribution and is slightly left shifted:

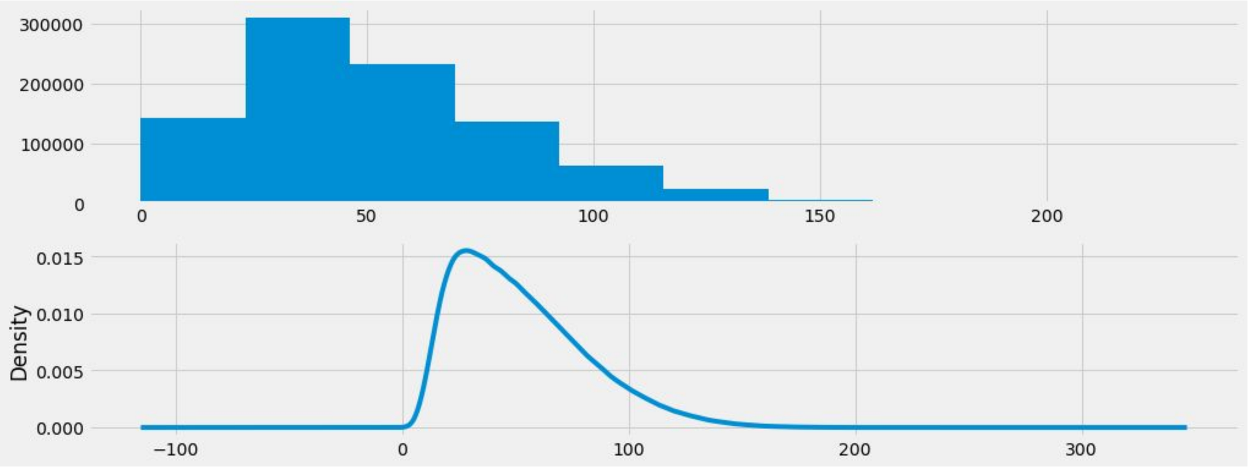


***Figure 1.1*** *Distribution and density of monthly total sales numbers from 2013 through 2017*

Median sales values have an upward trend over time. Total annual sales volume increased slightly each year, and there is also an increase in the spread, or middle 50% of the data, over time.

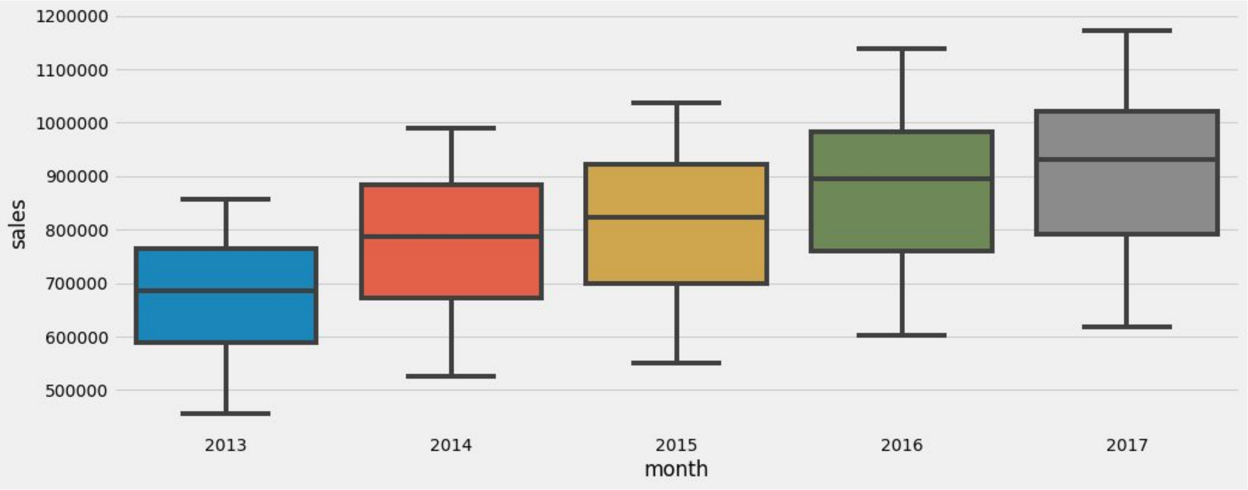


***Figure 1.2*** *Median and spread of item sales at all stores from 2013 through 2017*

Each year, total item sales climbs to a peak in July from January, and then goes back down to a trough with a slight peak in November. Reviewing the total sales numbers over time for all items in all stores reveals a time series with yearly seasonality and an upwards trend:

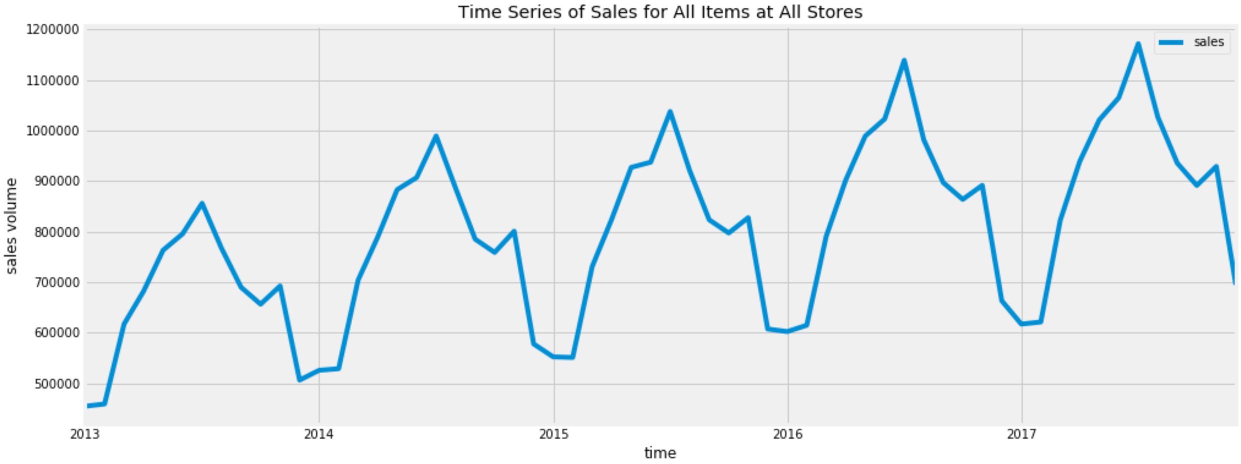


***Figure 1.3*** *Time Series of sales totals for all items at all stores from 2013 through 2017*

## 2.2 Individual Sales

For item 1 at store 1, daily sales numbers range from 4 to 50 with a mean of 20. The majority, or middle 50%, of daily sales numbers fall between 15 and 24, with the least number of observations below 8 and above 35. Unlike overall sales, density of individual item sales observations at one store shows the data is normally distributed:

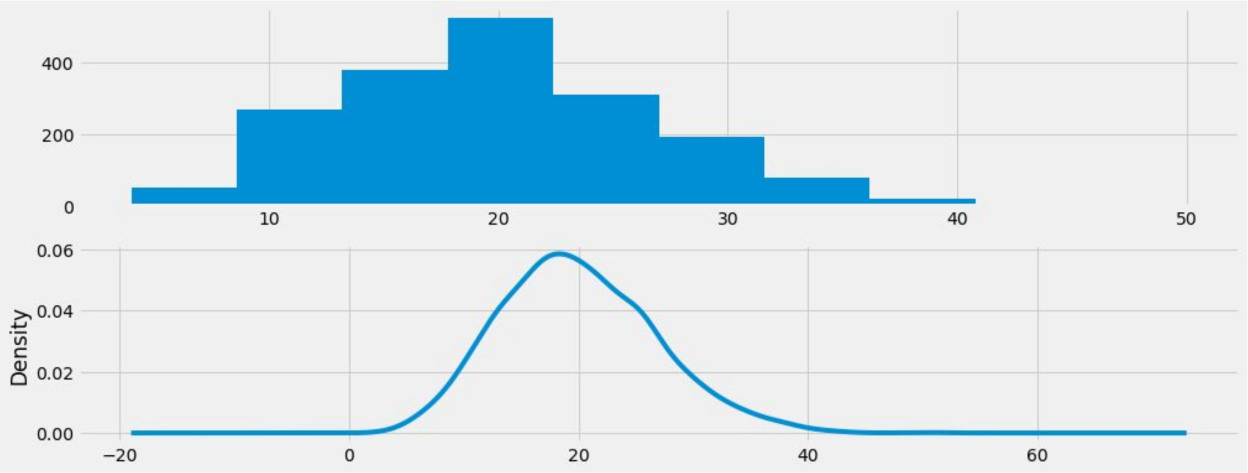


***Figure 1.7*** *Distribution and density of item 1 sales numbers at store 1 from 2013 through 2017*

Median values of item 1 sales increased from 2013 to 2016 with a slight decrease in 2017. While the total annual item 1 sales volume did not increase steadily over time, the spread or middle 50% of the data increased steadily over time:

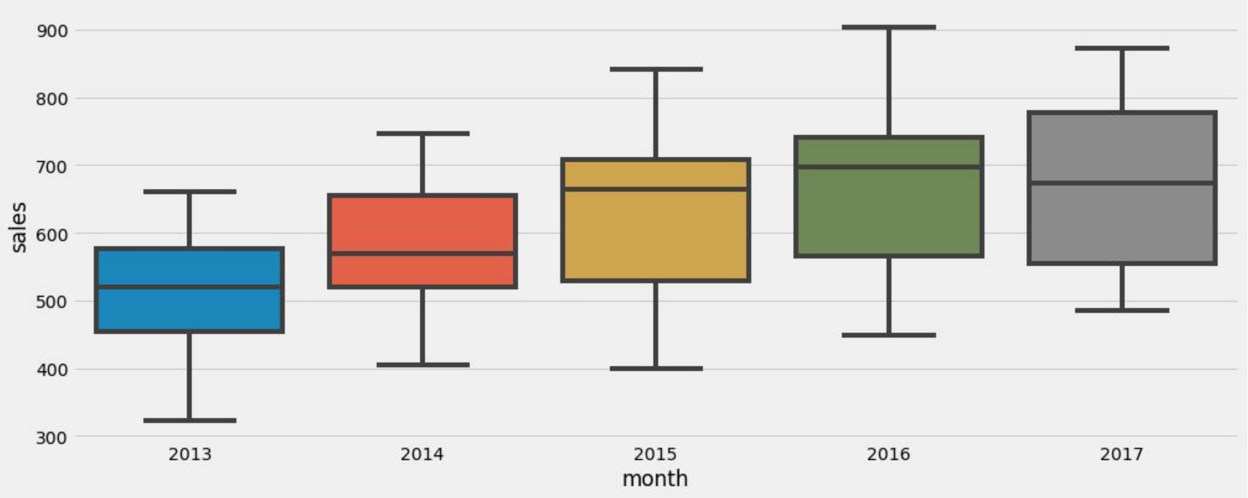
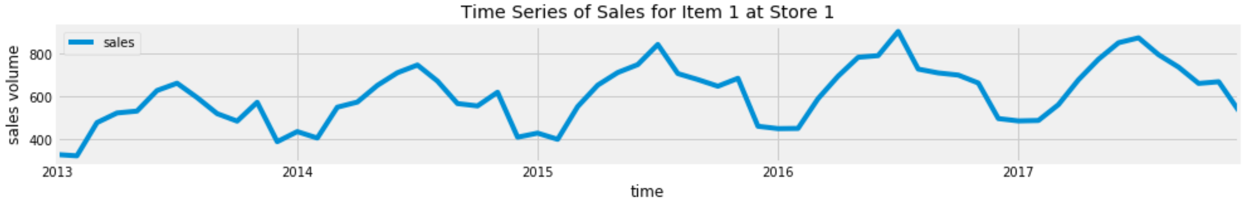
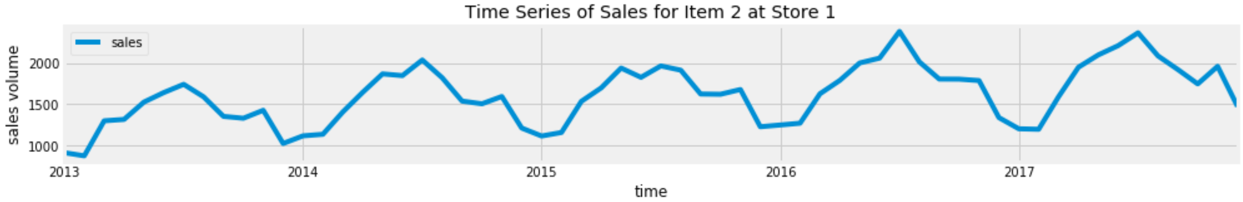


***Figure 1.8*** *Median and spread of item 1 sales at store 1 from 2013 through 2017*

For both item 1 and item 2 at store 1, the peaks and troughs are more irregular throughout the years. The trends and seasonality of the time series are less pronounced.


Time Series of Sales for Item 1 at Store 1


Time Series of Sales for Item 2 at Store 1

Decomposition of item 1 sales at store 1 confirms an upward trend and strong seasonality from year to year with peaks occurring in July of each year:

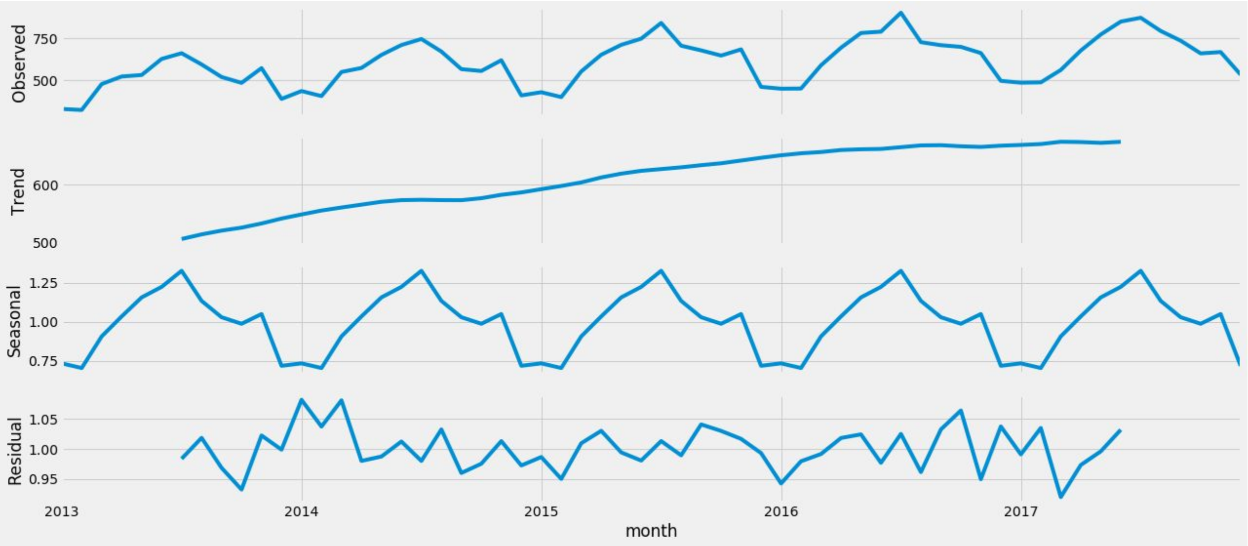


***Figure 1.9*** *Decomposition of item 1 sales at store 1 from 2013 through 2017*

# 3. Modeling Preparations

## 3.1 Extending Date Features

Combining test and training data, the dataset is extended with date features that describe each day with the week, day of year, month, etc. Figuring out how these features impact item sales is a step in preparation for fine tuning the time series models.

Box cox transformation is applied on the sales values to obtain normal distribution, which is required for training with the LightGBM framework. True and False values in the data are converted to 1 and 0 before training. Tree based learning algorithms from the LightGBM framework are used to determine feature importances.

Plotting the resulting feature importances shows the week, item, and day as the top 3 most important features having the biggest impact on sales numbers. Conversely, the store, day of year, and month are among the least important features:

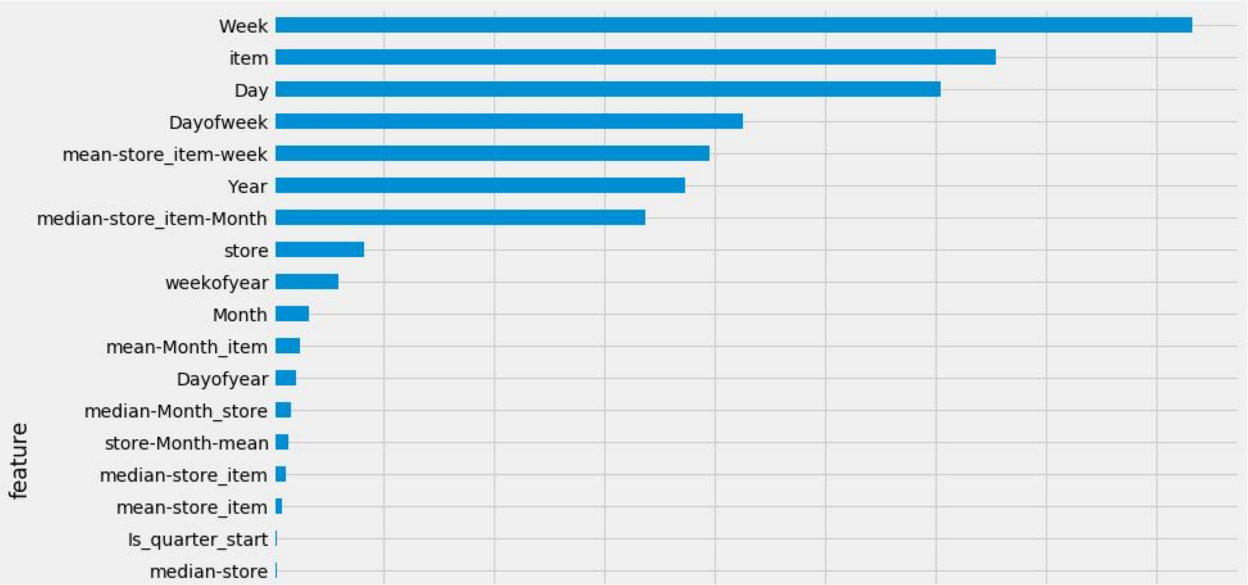


***Figure 2.1*** *Date features and their importances*

## 3.2 Data Transformation

In order to model a time series, it must be stationary. Most non-stationary series can be transformed into stationary series. Most time series models work with stationary time series because non-stationarity increases unpredictability. Newer models like Facebook Prophet provide higher flexibility in dealing with non-stationary time series. Stationary means that the distribution of the data doesn't change with time. For a time series to be stationary, it must fulfill 3 criteria:

- Zero trend: the series does not grow or shrink over time
- Constant variance: average distance of the data points from the zero line does not change
- Constant autocorrelation: how each value in the series is related to its neighbors stays the same

Identifying whether a time series is stationary or non-stationary is very important. If it is stationary, then we can use ARMA (autoregressive moving average) models to predict the next values of the time series. If it is non-stationary, then models like Facebook Prophet should be used.

Stationarity of an individual time series (for item 1 at store 1)  is checked below using the augmented Dickey-Fuller Test. Here the null hypothesis is that the time series is non-stationary. The test results comprise of a Test Statistic and some Critical Values at different confidence levels. If the ‘Test Statistic’ is less than the ‘Critical Value’, then the null hypothesis is rejected and the time series is stationary.

The results for the time series of item 1 at store 1 show the test statistic is -3.16, which is less than the 5% critical value but greater than the 1% critical value, and the p-value 0.02 is greater than 0.05. Since the series has a strong upward trend, we will use the strictest 1% critical value to test the hypothesis. Therefore, we accept the null hypothesis and the time series of item 1 at store 1 is non-stationary. Transformations of individual time series are necessary prior to modeling using ARMA models.

```
Results of Dickey-Fuller Test:
Test Statistic                   -3.157671
p-value                           0.022569
#Lags Used                       23.000000
Number of Observations Used    1802.000000
Critical Value (1%)              -3.433984
Critical Value (5%)              -2.863145
Critical Value (10%)             -2.567625
```

Differencing is the technique used to transform the individual time series for stationarity. In this technique, we take the difference of the observation at a particular instant with that at the previous instant. After first order differencing using Pandas, the test statistic is less than the 1% critical value and the p-value is much smaller than 0.05. We can now reject the null hypothesis and state that the transformed time series is stationary. The series also shows constant variance over time upon plotting its rolling statistics:

```
Results of Dickey-Fuller Test:
Test Statistic                 -1.267679e+01
p-value                         1.210928e-23
#Lags Used                      2.200000e+01
Number of Observations Used     1.802000e+03
Critical Value (1%)            -3.433984e+00
Critical Value (5%)            -2.863145e+00
Critical Value (10%)           -2.567625e+00
```

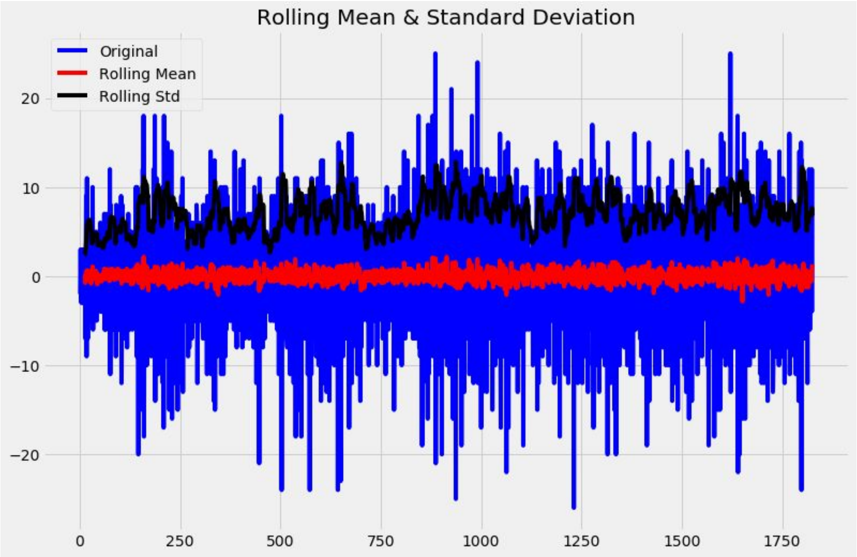


***Figure 2.2*** *Rolling statistics of item 1 sales at store 1*