Time Series Forecast with ARIMA Model and XGBoost

**Introduction**

Recently, I was approached by a food service company to conduct sales forecast on 2 restaurants they own. The data is anonymized, so I am able to share it here. The problem is time series, so I thought of writing a blog post to cover how one can use time series model, such as ARIMA, and tree model, such as XGBoost, to model a time series problem. In this blog post, I will cover the following:

1. Data Exploratory / Data Cleaning
2. Statistical Testing
3. Model Building
4. Forecasting
5. Improvement with XGBoost

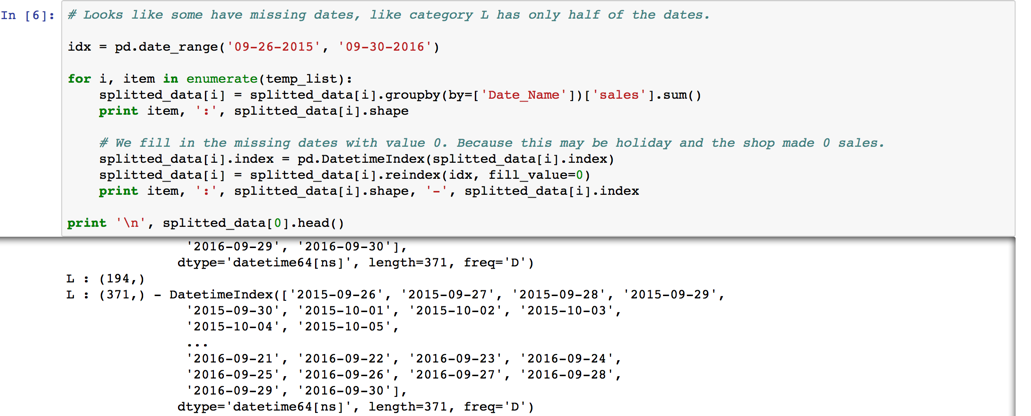
Without further ado, let’s get started with looking at the dataset.

**Data Exploratory / Data Cleaning**

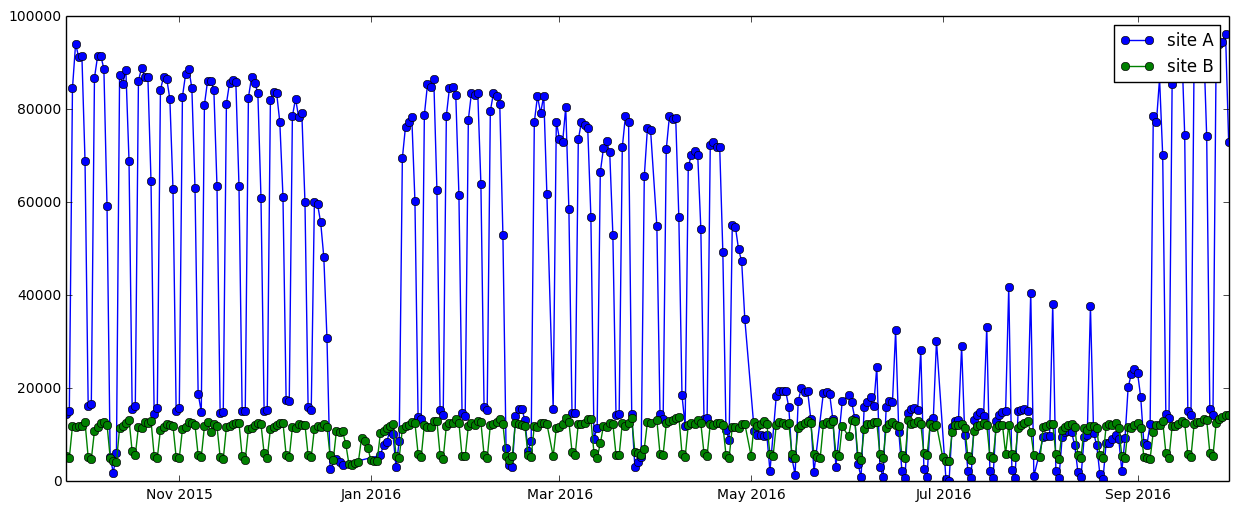
Taking a first glance of the data, we find our target variable that we want to predict, sales.



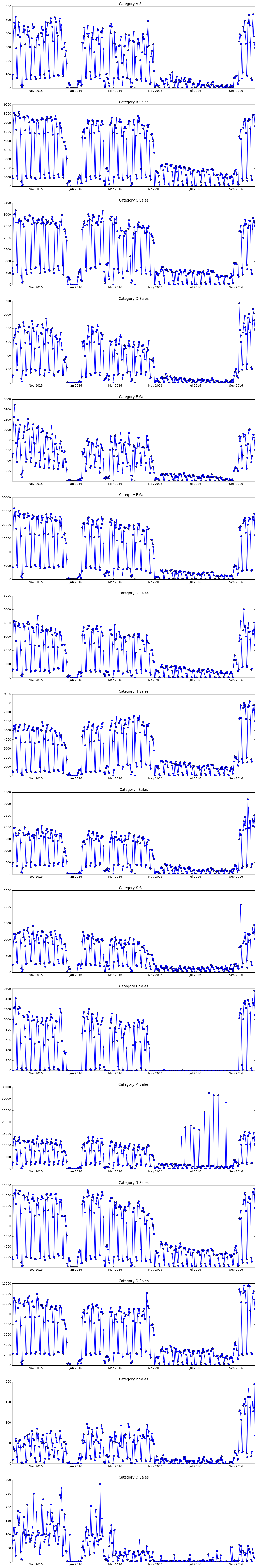
Our available features are, Date\_Name, subcategory, and Purchase\_Hour. However, I was instructed to forecast 90 days ahead, so hourly sales data becomes unnecessary and it is actually quite noisy if we keep it. So I folded the hourly sales data into daily data, i.e. combine Purchase\_Hour and Date\_Name. I’ve also filled in missing date, as there are some holiday date missing such as Christmas. In addition, the company has requested to forecast the sales on each individual category in addition to the sales of the whole restaurant. Thus, I needed to create a dataframe for each category.



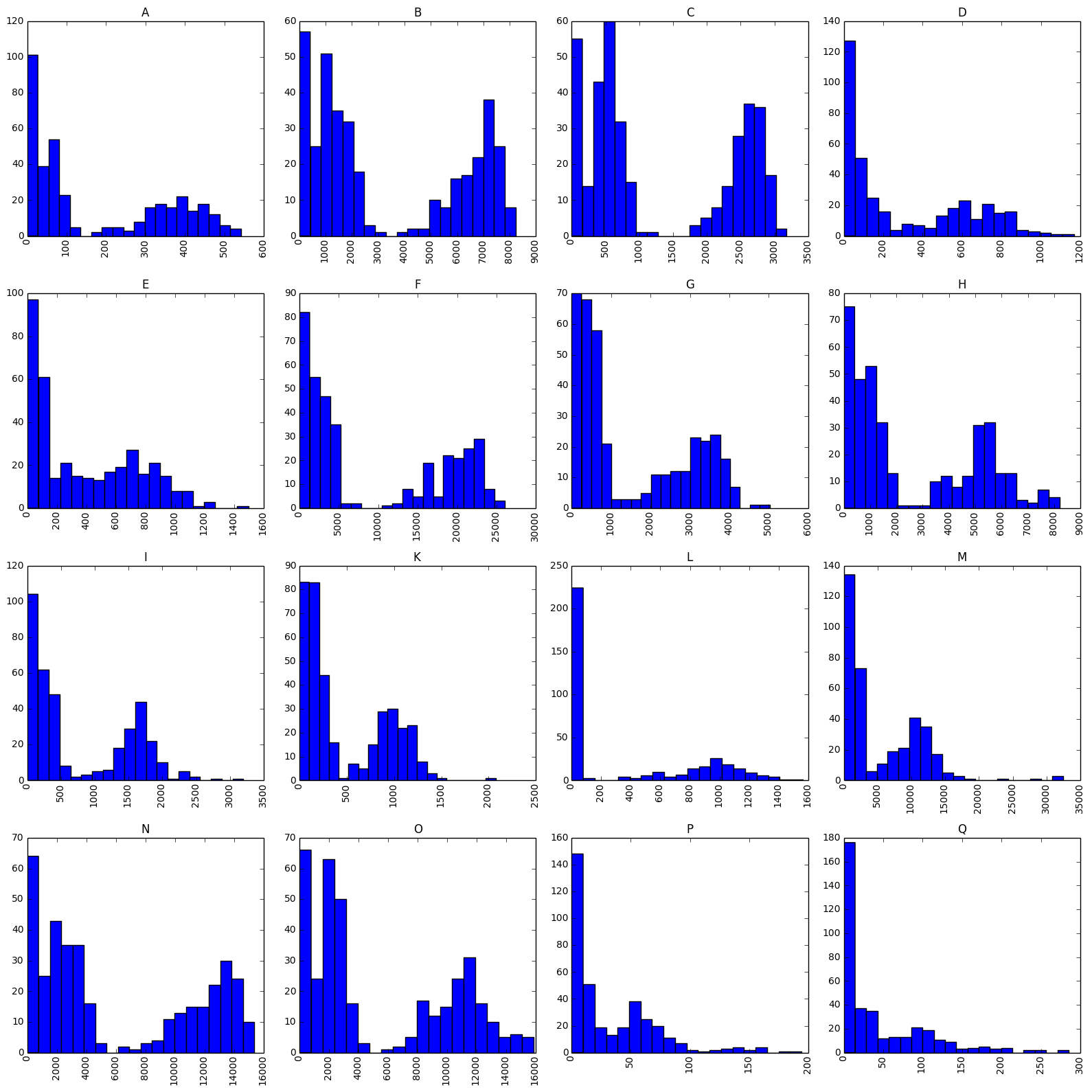
Now that we have the date filled, and the proper dataframe for each category, let’s make a time series plot to see the overall curve dynamics. Below shows the time series plot for the entire restaurant for 2 locations, site A and site B. We can see during Christmas, site A has no sales data, and during summer site A has much lower sales than usual. Both site shows a weekly seasonal effect that weekends have lower sales than weekdays.



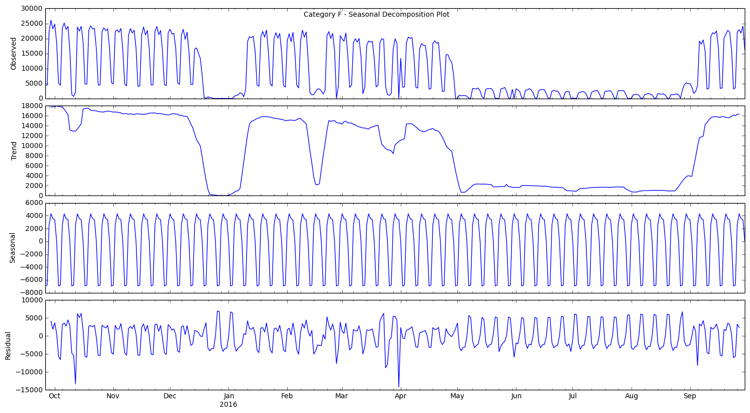
If we look at a specific category, for example category A and B, the category generally follows the overall curve dynamics:



Looking at the histogram plot, we find that we have 2 main groups sales. High sales during weekdays, and low sales during weekends + summer.



To further confirm the weekly seasonality effect, I plotted the Seasonal Decomposition plot for Site A. In the below we can definitely observe a seasonality effect on 3rd plot.



**Statistical Testing**

Now that we’ve confirmed the weekly seasonality effect, we need to conduct statistical testing to see if our dataset follows the ARIMA model’s assumption, zero mean constant variance. If not, then we will need to do mathematical transformation, such as log, to transform our data set.

I used the Dickey-Fuller Test, with 95% confidence interval, to see if our dataset passes the test.