Analysis of Sales at the TKTS Booths in New York City

Jonathan Calindas

May 3, 2017

Springboard Data Science Intensive

Capstone Project

Background: The Theatre Development Fund (TDF) operates the tourist landmark TKTS Booths in New York City that sell same day tickets to Broadway and Off Broadway shows for up to a 50% discount. There are four TKTS booth location throughout the city. The busiest location is Times Square, at the heart of the theatre district. Other locations are the South Street Seaport, Brooklyn, and Lincoln Center.

A mobile app has been developed for iOS and Android devices that will allow the user to see the shows that are listed at the booths without having to be there in person (they will still have to purchase tickets in person). The mobile app also serves as a comprehensive index of all the shows currently playing in New York City, including shows that are not on sale at the booth.

The Problem: Sales at the booths are subject to some fluctuation, but TDF management does not always know the cause of rises and dips in sales. Management needs a better model for sales that will allow them to better forecast sales on a weekly basis and detect when an anomalous event occurs.

Study Parameters:

1. Study dates: 1/1/2016 to 12/31/2016
2. To minimize the complexity of the natural cycle of activity within one week, we will examine the data by week.

Datasets to be studied and compared:

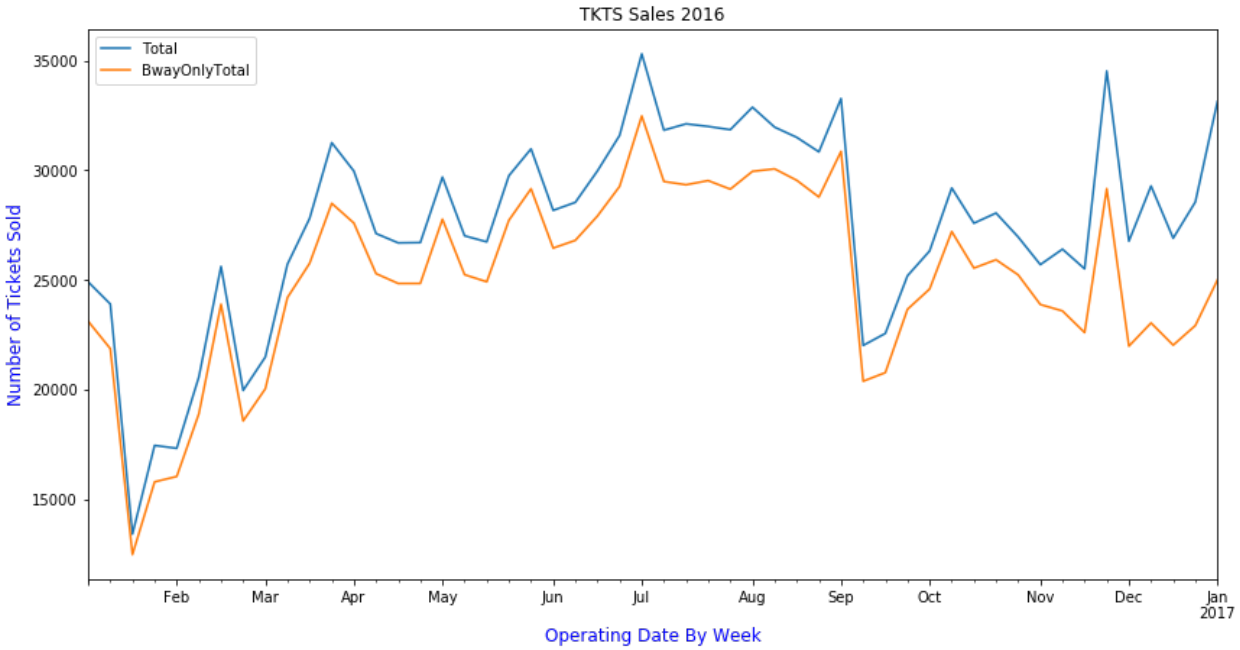
1. Tickets sales at each of the booths by week and by show.
    a. Sales figures are maintained weekly (not daily) and are kept in individual spreadsheets for each week of operation. Data had to manually be compiled into one spreadsheet that could be imported as a whole table.
2. Ticket sales for each show for all of Broadway as published by Variety.
3. Number of shows playing on Broadway
    a. Number sold at the booth. This is taken from sales data as well as from the software used by the sellers at the booth to post info to the app.
    b. Number of shows that sell at full price and don't appear at the booth. In addition to shows sold at the booth, a select number of shows sell very well and do not sell discounted tickets at the booth. A list of full show openings and closings are published at IBDB.com.

Much of the data will already be stored within TDF's databases.

Methodology: The bulk of the work will be to assemble and shape the data in a way that can be examined weekly. We will then plot these charts and analyze them to derive insights as to the behavior of the data. Finally, we will build an ARIMA time series model that can be used to forecast future sales.

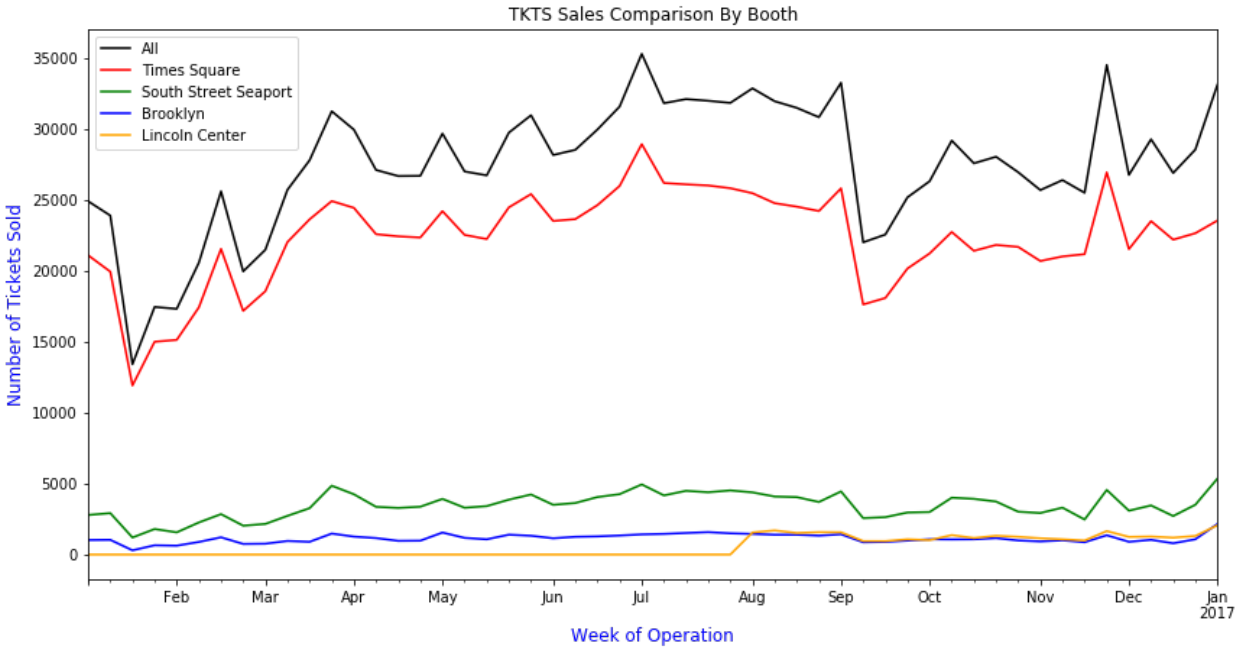
TKTS Sales

Before we examine the various factors that may influence the rise and fall of sales at the booth, it is important to establish a baseline with regards to the natural changes in booth sales.

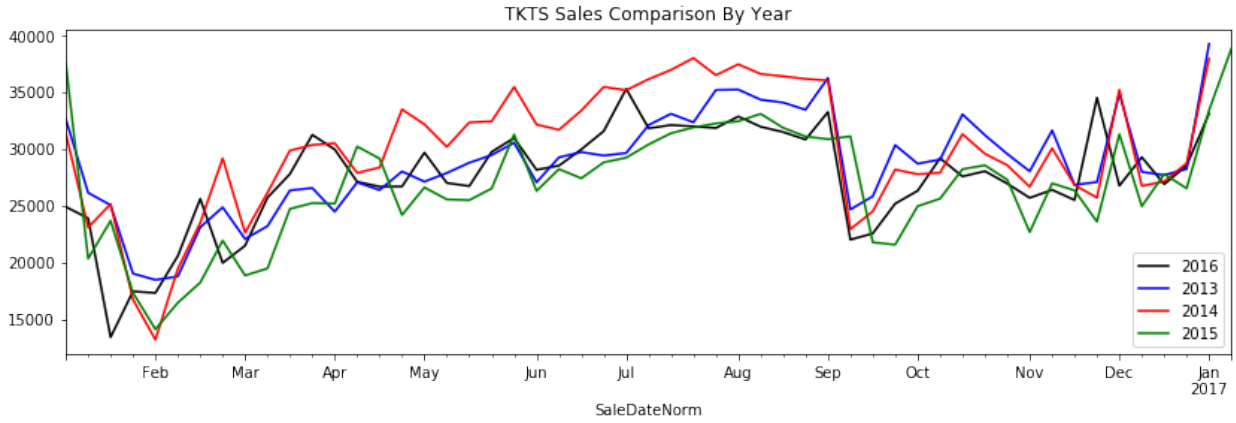


The blue line shows the total sales in all the booths for both Broadway and Off Broadway shows. The orange line shows only Broadway sales. We can see that Broadway represents the large percentage of shows sold at the booth and the gap between them only widens as more tickets are sold. The only exception to this is near the end of the year, in the months of November and December. This is attributed to the big show, Radio City Christmas Spectacular that is sold only at the end of the year. This is classified as an Off Broadway show, even though it is a big venue. Otherwise, we can also see the both lines dip and fall in the same way so factors that affect sales at the booth general affect both Broadway and Off.

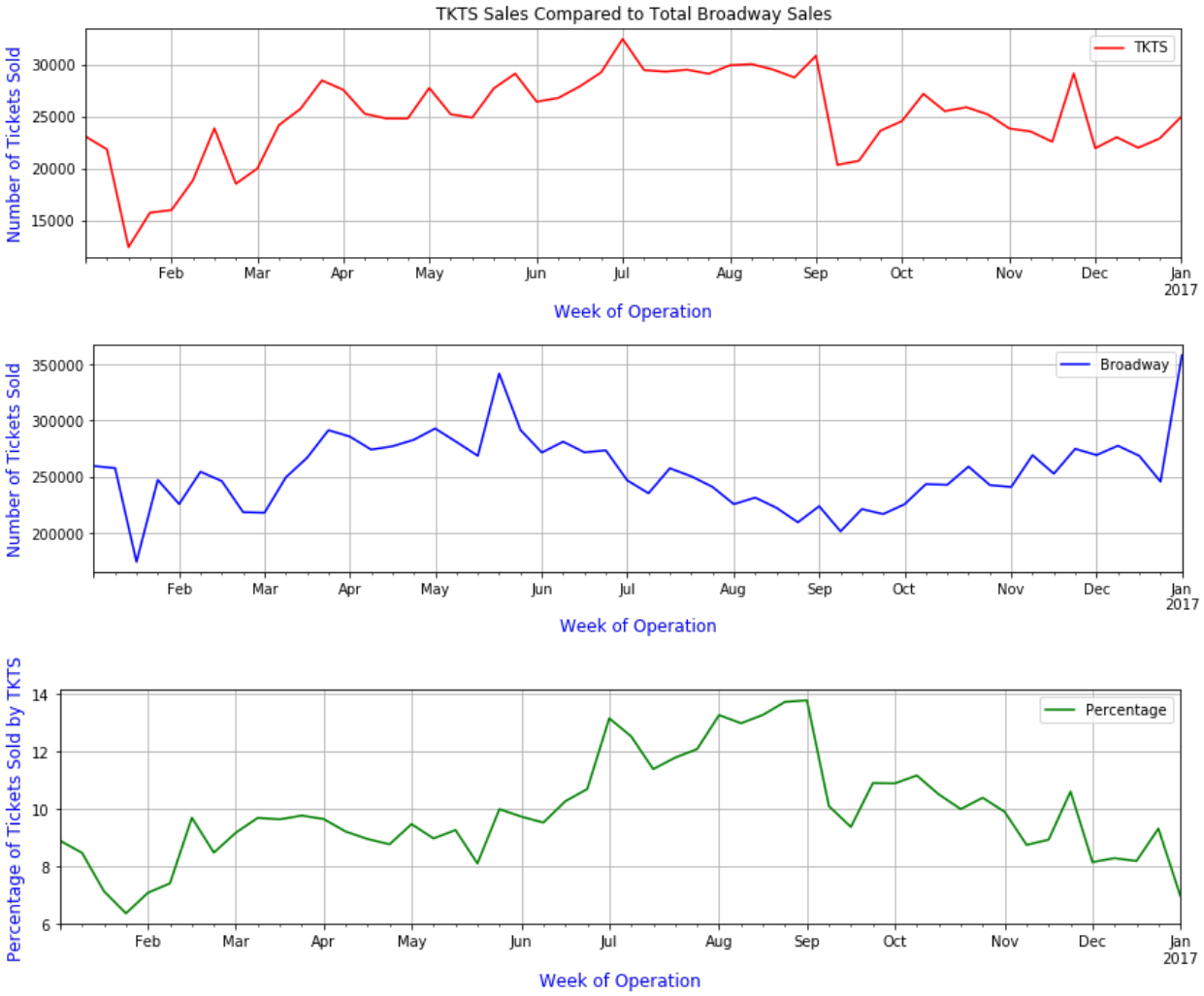
We can also see that the months of January and February and the worst months of the year with the summer months being the best months for the booth. We also see a precipitous drop in the first week of September as the school year begins. We also see great spikes in sales during holidays in the week of Valentine's Day, Easter and the week of spring break, mother's day, Memorial Day, the week of the Fourth of July, the week of Thanksgiving, and the week of Christmas.



The chart above shows ticket sales by booth. We can see here that the vast majority of ticket sales are sold at the Times Square location. Note that the Lincoln Center location did not open until late July, which is why it is zero up to that point. We can see that every location generally follows the rise and fall in sales proportionally in tandem, so all booths see the peaks and valleys at the same time. If it doesn't seem to rise as much in the chart above, it's only because the scale is much smaller.



When we compare the trends compared to previous years, we can better determine the natural rise and fall of sales throughout the year. The minor shifts are attributed to the migration of days as they fall within the week number. For example, in January 2016, week 1 ended in Jan 10, whereas in other years week 1 ended much earlier. We see that the lowest point in sales always occur in the January after the week of New Years. Then there is a gradual rise in the spring months, and peak sales in the summer months. There is always a precipitous drop in the first week of September, a rise in sales in the fall, and a sudden spike in sales during special holidays, such as Valentines Day (Feb 14), Easter, Mother's Day, Memorial Day, the week of Thanksgiving, and the week of Christmas and New Years. We can also see that not all holidays produce a spike in sales every year, for example, the Fourth of July produced a spike in 2016, but previous years were mostly flat.

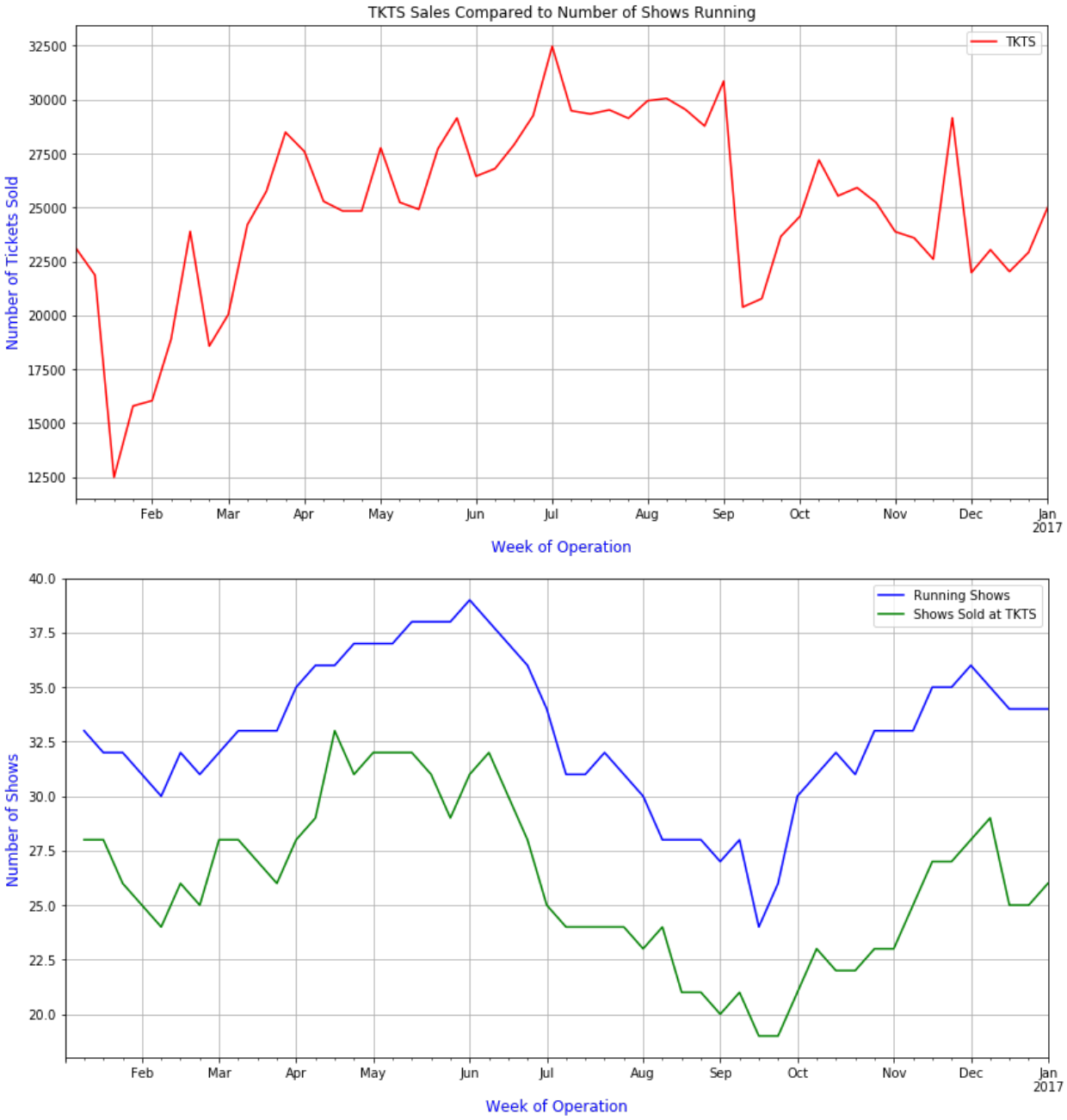


Now we will compare TKTS sales compared to all of Broadway. The first chart contains Broadway only sales for the booths. The second chart shows sales for all of Broadway, which includes full price tickets sales and sales through other ticket outlets. We need a different graph because the scale is much higher in the second chart. The third chart shows the percentage of total ticket sales sold through TKTS (market share).

We can see that Broadway sales peak during the spring and the fall months. This is generally when new shows open and close. We see a sharp rise in sales in the end of May to the beginning of June, which is attributed to the end of the Broadway season and the Tony Awards. Shows see a rise in sales due to the excitement of the Tony awards and it is the when Tony voters come to see the shows.

As we compare Broadway versus TKTS, we also see the rise in sales during the fall and spring months, but sales continue to rise into the summer for TKTS, while the rest of the industry falls during the summer. Both charts see the lowest point of sales in January, but Broadway is generally steady during holidays, with the exception of the Christmas/New Years holiday.

From the third chart, we see that TKTS has the highest market share during the summer months, with peaks during specific holidays.



In the charts above, we compare TKTS sales to the number of shows playing on Broadway in any given time. The blue line indicates the total number of shows, while the green line indicates the number of shows that are sold at the booths. The slow rises during the spring and fall months show the season as shows open during the fall and spring months. Shows try to stay open until the Tony awards in June, but see some shows close during the winter months of January and February, and a large number of shows close right after the Tonys. They begin to open again in the fall from September on.

When we compare the blue line to the green line, we generally see that about the same five or six blockbuster shows sell at full price and not at the booths, with the exception of a few shows that will sell really well at the start and then begin to sell at the booth after a while.

We see from the chart, however, that the number of shows generally do not affect the short term level of sales at the booths, although sales are generally high (along with the industry) during the spring and fall months. The booth still does well, however, with the shows that remain after the Tonys and the fewer number of shows still playing.

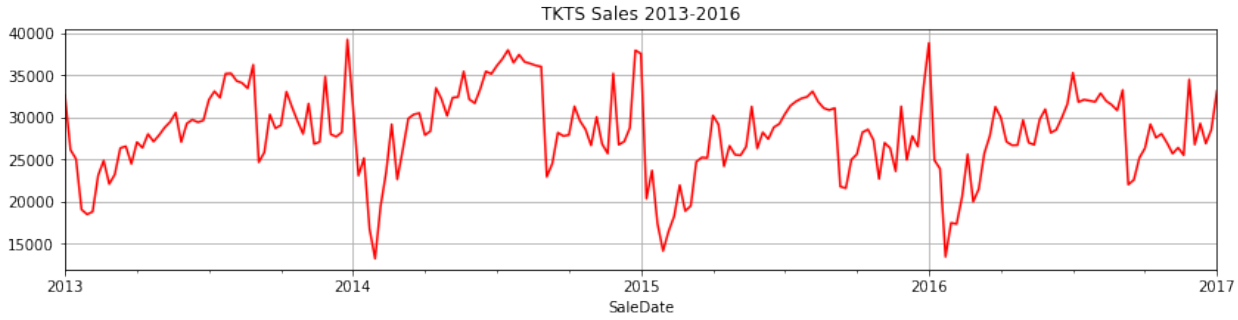
Other Notes About Sales Data

In compiling the sales figures above, I chose to only use the number of tickets sold rather than the actual dollar amounts in sales. This simplifies the data a great deal, but brings with it some important caveats. Some shows will sell at the booths a smaller discount percentage than other shows, so their dollar amounts may be higher even if they sell fewer tickets. Also, during peak Broadway season, Tony voters are usually comped, so even though tickets are high, actual dollar figures will probably be lower or remain even. Further investigations into these figures should yield more insight into the behavior of ticket buyers.

Creating a Model to Forecast Sales

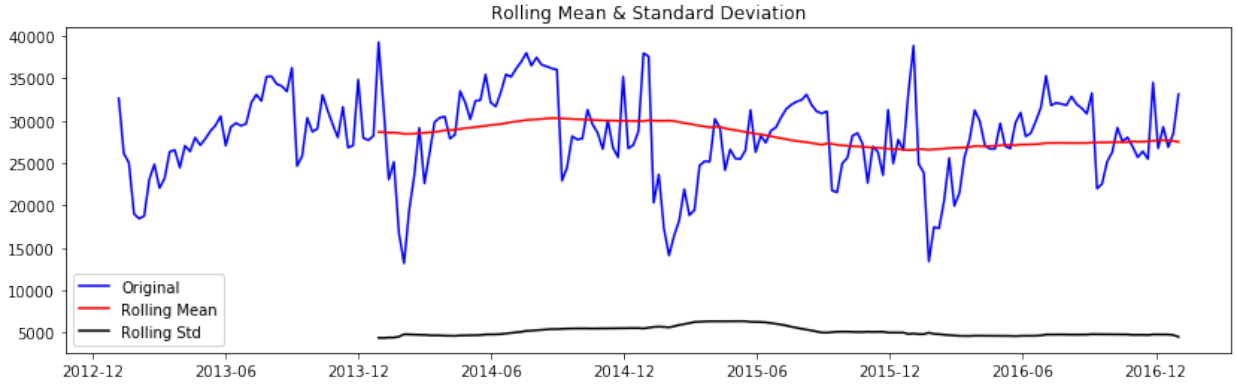
In this section, we will create an ARIMA model on the time series sales data and use this model to predict future sales. Our model will be built on sales from 2013 to 2016. We will then use this model to forecast the first half of 2017 and compare it to sales in 2017 to date.

To start, we first view a plot of the entire time series:



We can see here that each year has the same basic pattern, which is the seasonality of the time series. We can also see that some years have a higher mean than others, for example, 2014 has much higher sales, overall than 2015. This, however, does not show an overall increasing or decreasing trend, just a year by year fluctuation.

The main requirement of forecasting using the ARIMA model is that the time series be stationary, which means that the mean and variance must remain constant over time. To test this, we will plot a rolling mean and a rolling standard deviation. Using Pandas (Python Library), for each point in the time series, we will take the mean of the previous cycle (in this case, 52 weeks), and do the same with the standard deviation. Because we need the previous 52 weeks to compute the mean, the rolling mean plot starts at the first week of 2013.



We see that the rolling mean and standard deviation remain mostly flat with some slight trending.

Another test of stationarity is the Dickey-Fuller test. We will use the Statsmodels python library to compute the values for our time series.

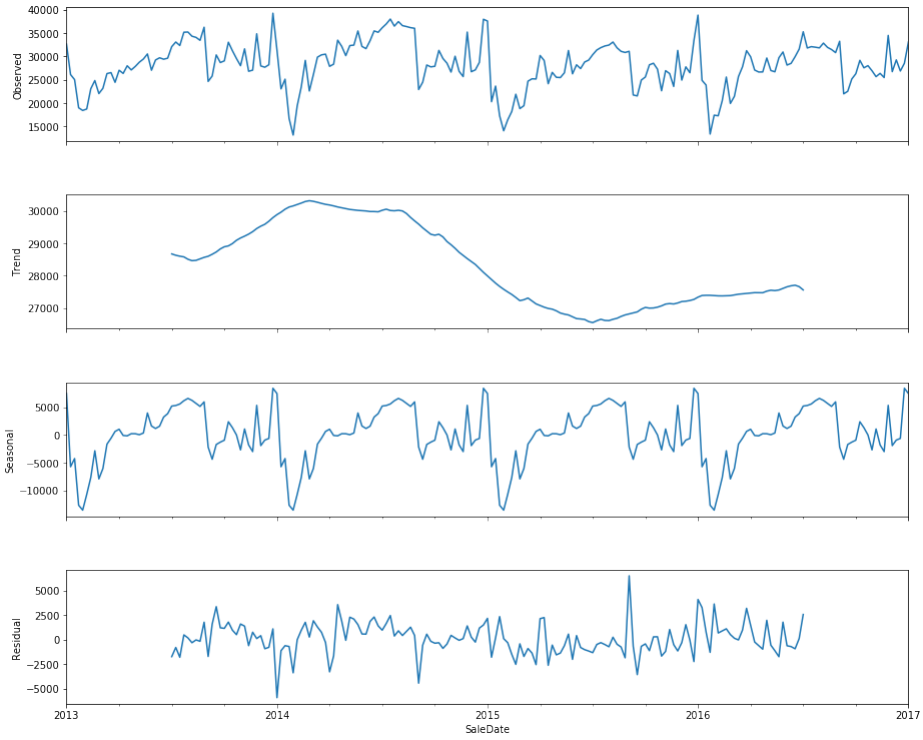
```
Results of Dickey-Fuller Test:
Test Statistic                  -3.431417
p-value                          0.009933
#Lags Used                       9.000000
Number of Observations Used    199.000000
Critical Value (1%)             -3.463645
Critical Value (5%)             -2.876176
Critical Value (10%)            -2.574572
dtype: float64
```

The Dickey-Fuller test uses hypothesis testing to test the stationarity of the model. If the Test Statistic is greater than the critical values (ignoring the negative sign) then the alternative hypothesis is correct and the series is stationary. In the results above, our Test Statistic is slightly below the 1% critical value, but is significantly higher than the 5% critical value. So we are almost 99% sure that the series is stationary.

Because the series is stationary, we will not need to account for trending or perform any differencing with our data to make it stationary.

Our next step is to decompose the time series into its various components. We will compute the trend, build a seasonality model, and find the residuals. The residuals will be what remain after the trend and seasonality are removed from the time series, and is what we will build the ARIMA model on.

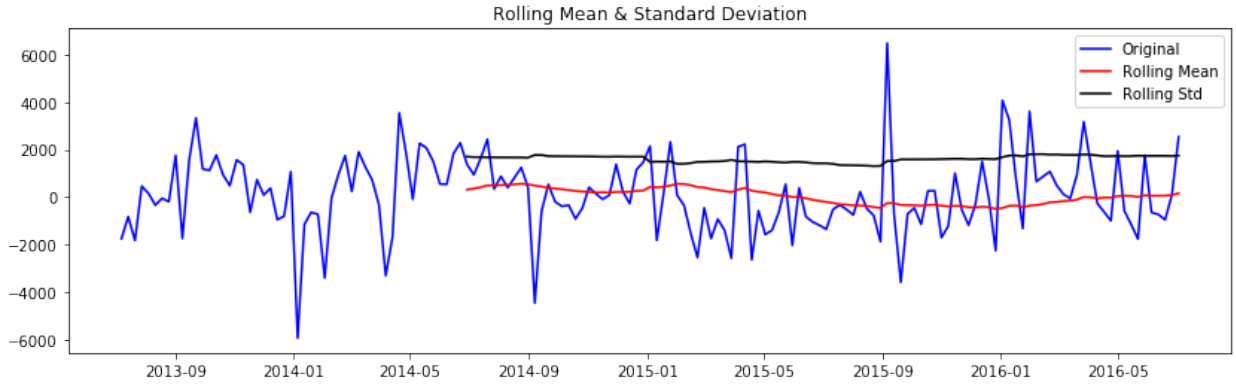
Again we will use the Statsmodels library to perform this computation:



Because trend is calculated using the 26 weeks before and after each data point, the first 26 weeks (half of a cycle) and the last 26 weeks of our time series is not calculated. Because of this, our residual data set also ends at the first half of 2016. We will use this "lost" half of 2016 towards our forecasted values comparison.

When we build our ARIMA model, we will create a forecast of one year, which will encompass the second half of 2016 and the first half of 2017.

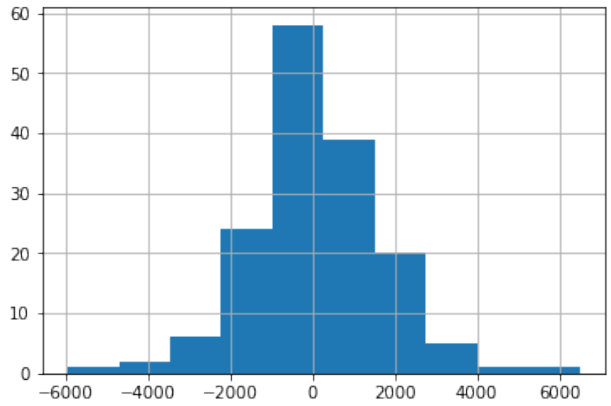
The residual data set at the end is what we will use to build our ARIMA model. First let us test the stationarity of this data set:




```
Results of Dickey-Fuller Test:
Test Statistic                  -4.188552
p-value                          0.000689
#Lags Used                       3.000000
Number of Observations Used    153.000000
Critical Value (1%)             -3.473830
Critical Value (5%)             -2.880623
Critical Value (10%)            -2.576945
dtype: float64
```

After removing seasonality and the trend, our residual data is even more stationary than our original time series.

We see also that our residuals have a mostly normal distribution:



To compute our ARIMA model, we will once again use Statsmodels. The function requires three parameters, the p parameter is the scale of the AR (auto-regressive) part of the computation, the q parameter is the scale of the MA (moving average) part of ARIMA, and the d parameter is how much differencing was required to make the model stationary. Since we did not use any differencing, our d parameter is 0. We will now play with the p and q parameters to give us the best AIC and BIC scores:

(p,d,q)	AIC	BIC
(1,0,0)	2678.757	2777.925
(2,0,0)	2679.886	2782.111
(3,0,0)	2764.929	2780.210
(4,0,0)	2760.633	2778.971
(5,0,0)	2762.616	2784.010
(6,0,0)	2764.551	2789.001
(0,0,3)	2769.273	2784.554
(0,0,4)	2760.789	2779.127
(0,0,5)	2762.469	2783.863
(0,0,6)	2764.392	2788.842
(2,0,1)	2770.646	2785.927
(1,0,1)	2768.788	2781.013
(5,0,3)	2766.950	2797.512

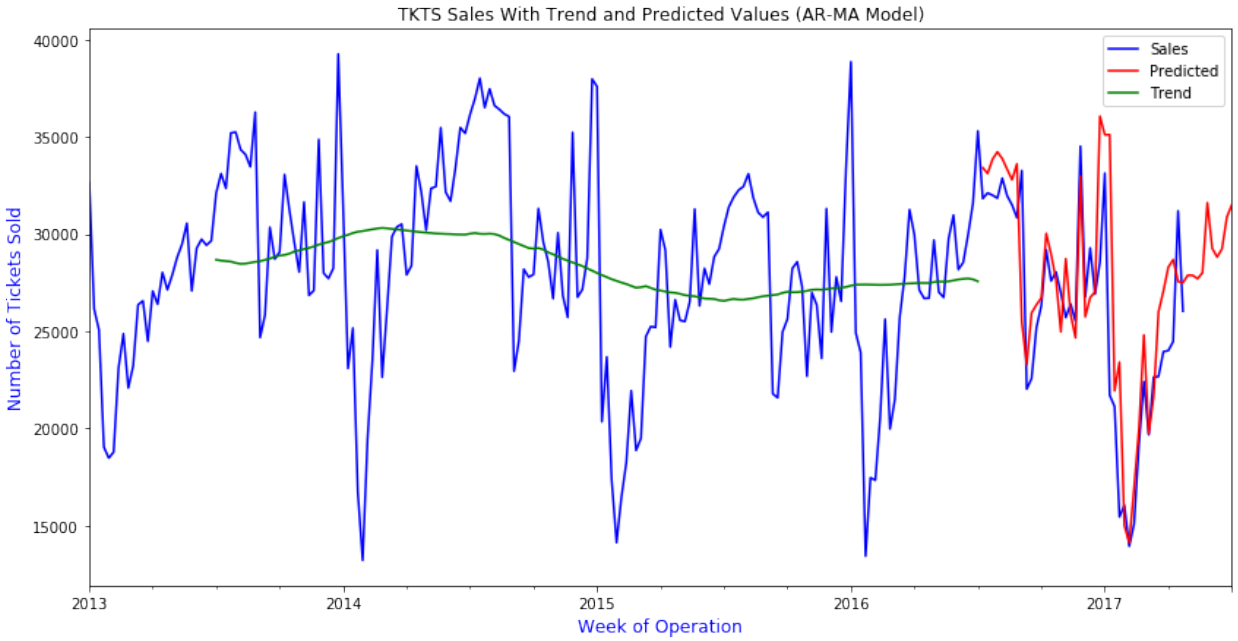
We will use the three sets of values (in red) to compute an ARIMA model and build a forecast using that model. I've included the third set (2,0,1) because the computed coefficients and standard error are much lower than the (4,0,0) and (0,0,4) sets, even if the AIC and BIC are much lower.

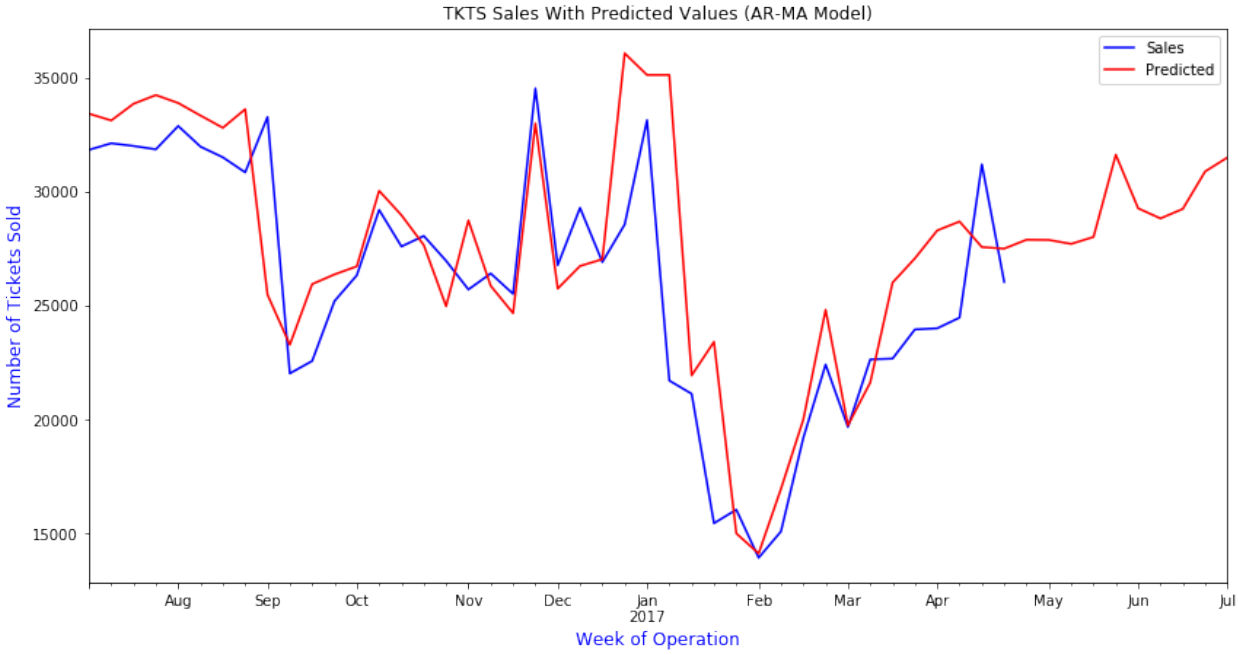
(p,d,q)	AIC	BIC	Std err	Coeff	z
(4,0,0)	2760.633	2778.971	215.307	82.34	0.382
(0,0,4)	2760.789	2779.127	179.413	88.91	0.496
(1,0,1)	2768.788	2781.013	147.796	95.72	0.648

After building our three ARIMA models, we then create a forecast of one year. As stated above, we will forecast the second half of 2016 and the first half of 2017. For each of our models, we will then compare to the actual sales for that period and compute the root mean squared error.

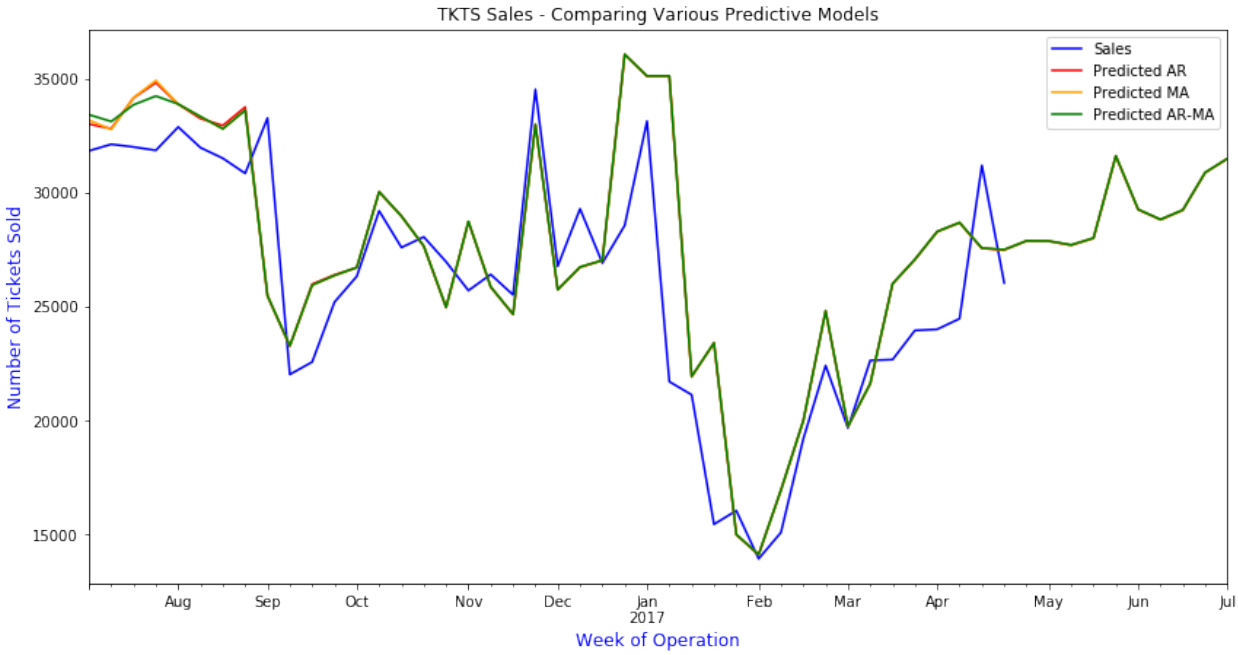
Model	Parameters	Root Mean Squared Error
AR-Model	(4,0,0)	3513.837
MA-Model	(0,0,4)	3513.828
Combined AR-MA Model	(1,0,1)	3503.468

We can see above that our combined AR-MA model performs the best with the lowest RMSE.





We can see that our model over predicts the Dec-Jan period quite at bit, as well as the Aug-Sept and March periods. However, if we look at our previous data, we see that sales in Dec 2016 to Jan 2017 were much lower than previous years, with the only period where sales beat predicted being the week of Easter 2017 (Apr 16).



Comparing our three ARIMA models, we can see that they are pretty much identical with some very slight variations from Aug to Sept.

Conclusions

Analyzing the sales from 2016 and from the past 3 years, we can build a model of behavior of sales at the TKTS booths from a year to year basis and can detect severe anomalies when they happen.

We know that sales at the booth follow a seasonality trend every year and that sales spike around holidays, such as Valentines Day, Easter, and Thanksgiving.

Sales at the booth are at their highest during the summer and at the month of December.

We can see that the general trend up or down between years is about 3,000 tickets, at least based our data. We can further generalize this trend if we go back several more years.

Our root mean squared error on our model states that we will be off on our predictions an average of about 3,500 tickets. But since our sales can go from 13,000 to 39,000, this is a fairly accurate model.