Assignment 2
Exploratory Data Analysis of Broadway performances

For this assignment, I chose to work on the Broadway dataset made available by the Broadway League. Analyzing this dataset could be of interest to theatre operators, producers, and the audience.

Source: https://corgis-edu.github.io/corgis/csv/broadway/

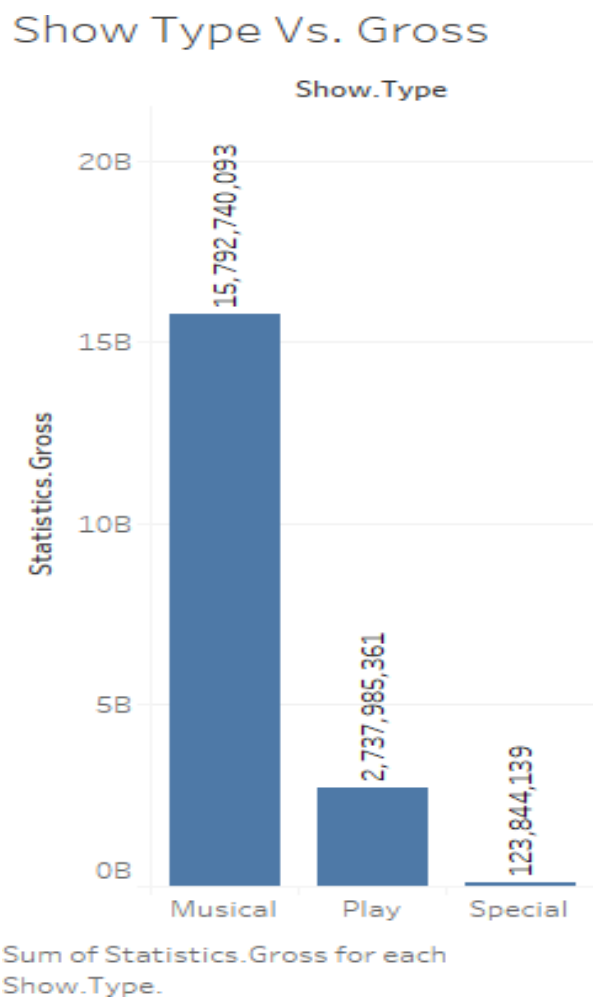
This version of dataset was provided by Austin Cory Bart (acbart@vt.edu).

The dataset is of CSV type and has been cleaned with the data interpreter tool of Tableau. The data has been collected between 1990 and 2006 through each week. All the visualizations in this analysis have been done using Tableau.

Broadway refers to the theatrical performances presented in theatres located along Broadway in New York City. Since, there are three main types of performances presented at Broadway, it is intriguing to know which show type makes the highest yield.

Which type of performances made the highest gross yield?

The “Gross gross” shows the total amount made by the performances. From this visualization, we can see that musical performances were highly preferred in comparison to plays and special performances. The bar chart adopts position encoding and it makes it easier to estimate that musical performances were nearly 6 times more profitable than the plays.



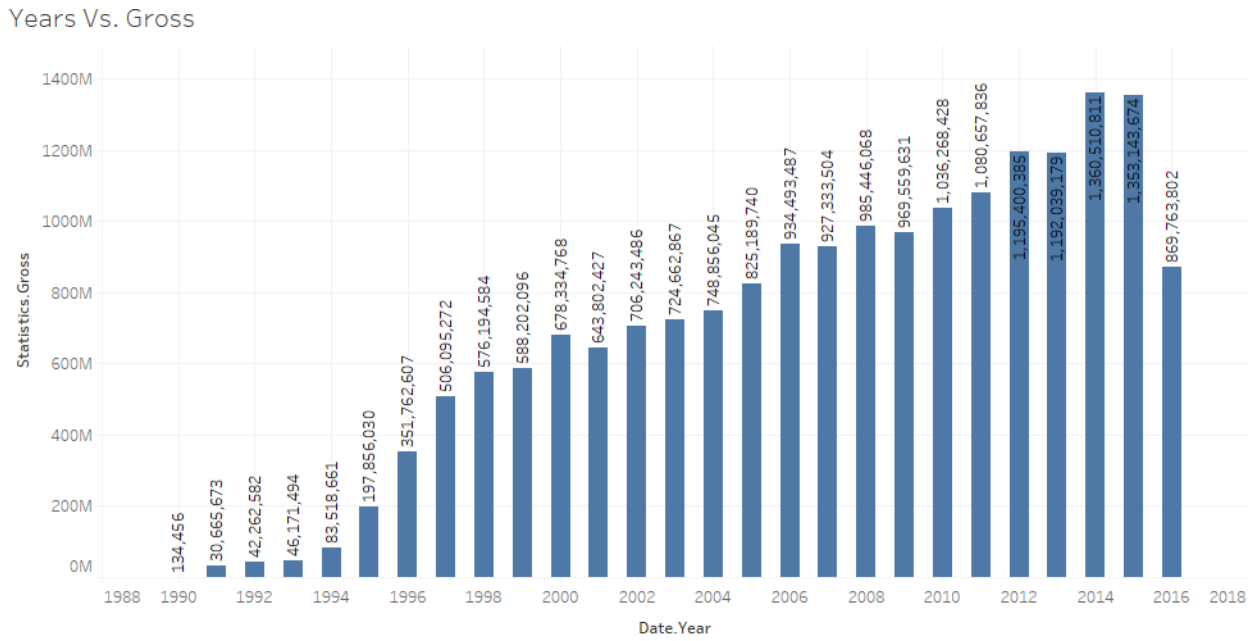
Sum of Statistics.Gross for each Show.Type.

Now that we know that the audience preferred musical performances, we can see when the gross yield was highest.

Which year made the highest gross yield?

Since we have the data between 1990 and 2016, it is interesting to see the trend in the gross yield over the years. From the below chart, we see a gradual increase over the years with a few fluctuations in between. It is obvious to note that the year 2014 has been the highest grossing year in this time period of Broadway Theatres.

We also note that unlike a minor fluctuation, the year 2016 has a drastic drop in the yields and is close to 2005 levels. It gives us an insight that some factor has played a role and it would be ideal to analyze and address that. If the number of performances hosted was low and in turn the audience attendance was low, then the resulting gross could be low. In that case, it could be due to lower availability of actors and more analysis on why low performances were hosted should be focused. However, if the number of performances had increased or remained the same and yet the audience turnout was low, then it will be helpful to focus more on audience experience and satisfaction.



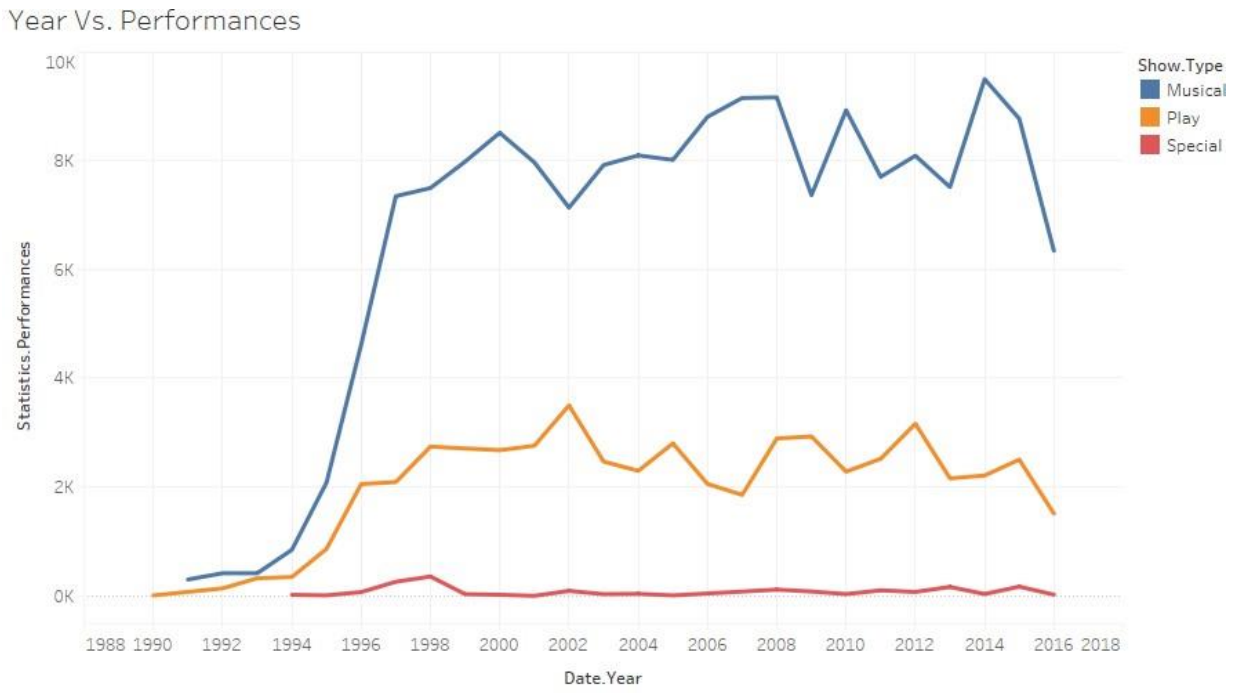
The plot of sum of Statistics.Gross for Date.Year.

It will now be effective to see if the gross yielded over the years are in tune with the number of performances hosted.

Trend of performances hosted over the years:

In tune with the previous charts, we see that the number of musical performances were higher in general over the years. This explains why they yield the highest gross. This chart shows that initially in 1990, only play performances existed. Between 1992 and 1996, musical performances became popular and had a high increase in the number of performances. Though relatively low, play performances too had shown an increase. Special performances were introduced from 1994 and had gained a minor popularity in 1998 but from then on, they had largely been performed at the same levels. It is also seen from this chart that, there has been a drop in the number of performances in the year 2016.

This could be because of non-availability of data over the entire year 2016 as it is the latest year in the dataset. This supports why the gross as well as number of performances have been low.



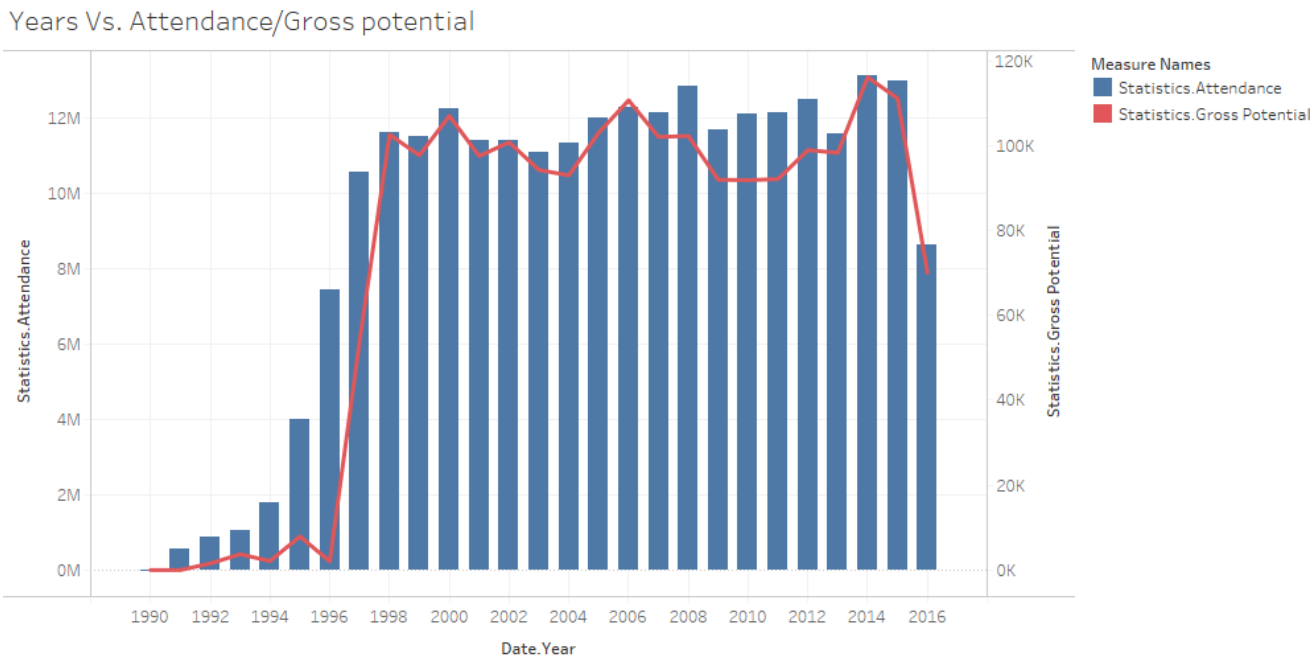
The trend of sum of Statistics.Performances for Date.Year. Color shows details about Show.Type.

Has gross potential been an effective reflection of audience attendance?

The Gross Potential is defined as the maximum amount that can possibly be earned considering the ticket prices, seating capacity, and the number of performances. This number indicates what could have been achieved and it is reported as zero when there is an absence of such predicted data.

It is intuitive that though the attendance and gross potential vary in scale, they must be correlated. It is seen that gross potential has not been calculated in 1990,1994 and 1996. Post 1998, the calculations have been well formulated and is very reflective of the audience attendance.

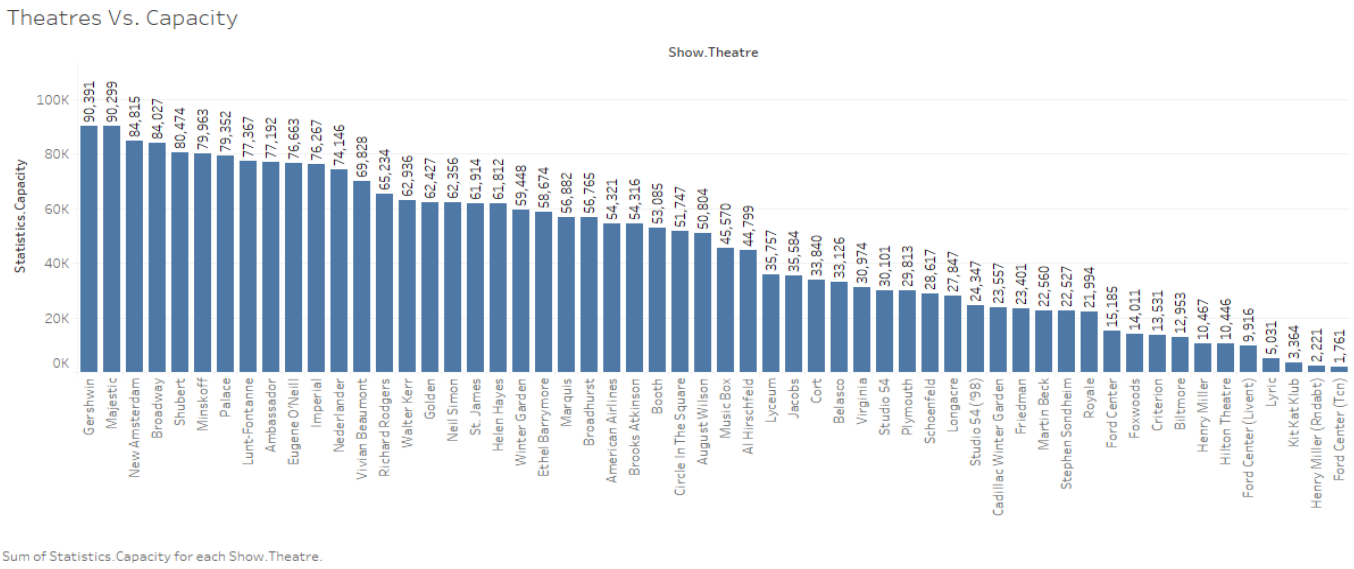
Between 2003 and 2004, increase in attendance is seen but a decreasing gross potential is predicted. This could be due to a drop in the ticket prices or more than expected audience attendance. Similarly, the stagnant gross potential between 2007 to 2008 and 2009 to 2010 while there has been a visible increase in attendance could indicate the changes in ticket prices (could have been cheaper than expected) or more attendance (more than expected).



As an audience, I would be more interested in knowing more about the theatres running the show. This leads us to analyze the following questions.

Where am I more likely to get a ticket?

Arranging the capacity of theatres in descending order indicates which theatre has the maximum capacity. It indicates that I am more likely to get a ticket at Gershwin followed by Majestic and so on. As an audience this data is more useful.



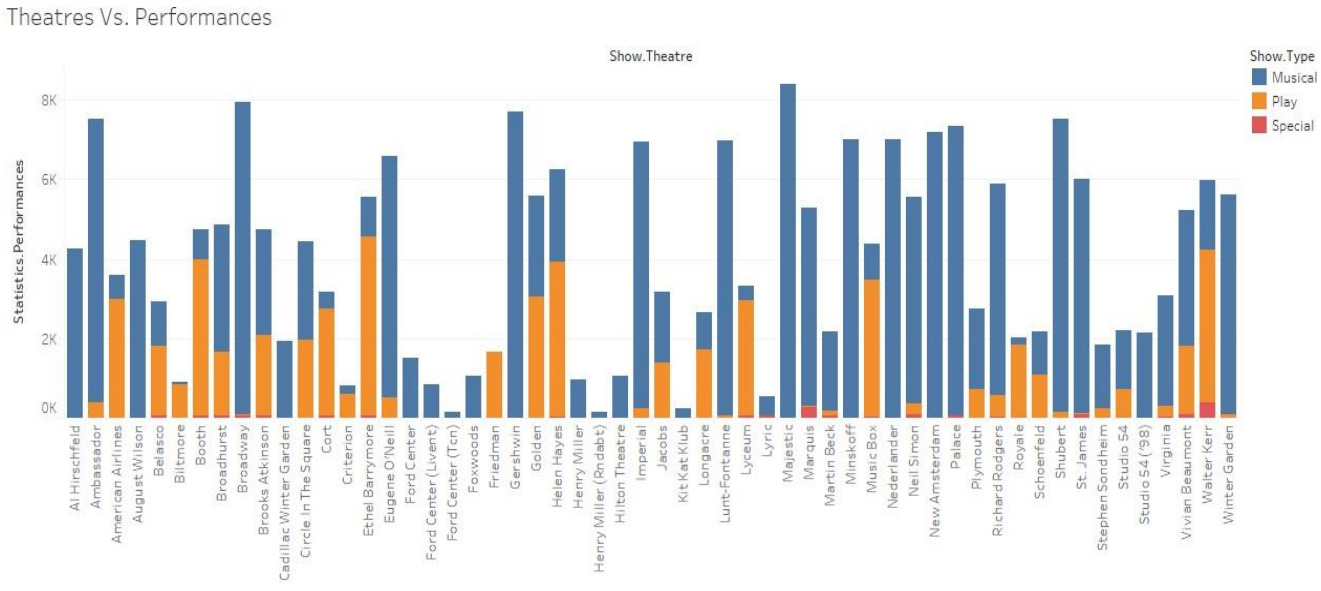
Sum of Statistics.Capacity for each Show.Theatre.

But I want to watch a play rather than a musical.

Which theatre should I check?

As an audience, if I am more worried about the type of performance rather than its capacity, then the following chart could be more indicative. I will know that the chances of watching a musical is highest at Majestic. For plays, it is Ethel Barrymore and for special performances, it is Walter Kerr. Since these host the maximum number of performances in that category.

Using a length chart here instead of position bar chart reduces overloading of data and makes it easier to interpret.

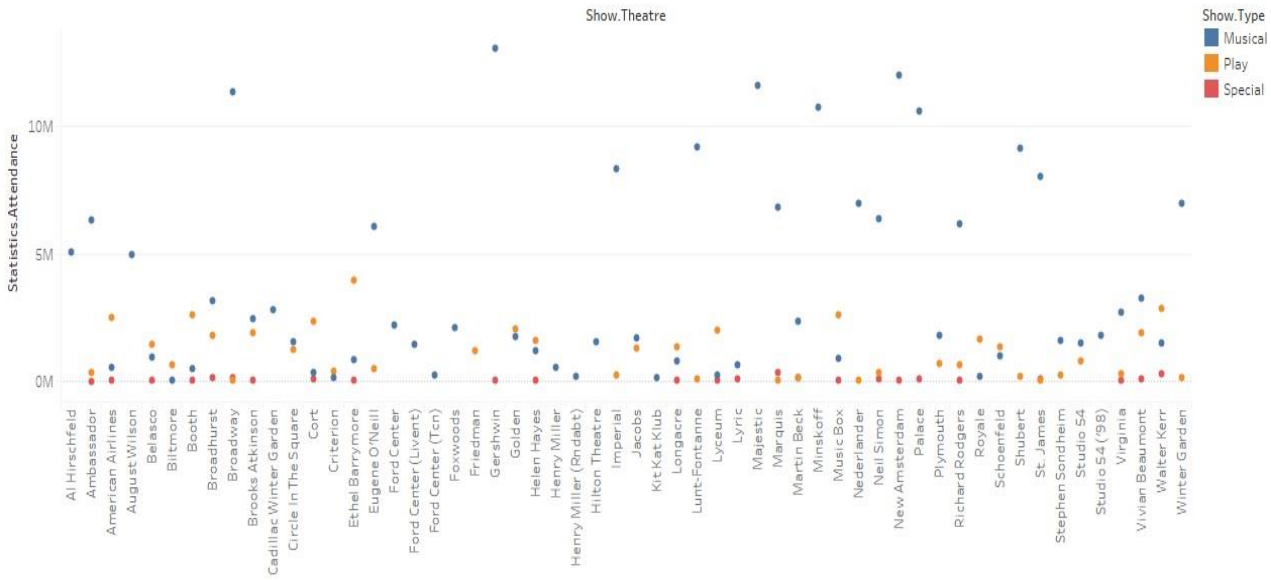


Sum of Statistics.Performances for each Show.Theatre. Color shows details about Show.Type.

Does more number of performances and capacity indicate more attendance?

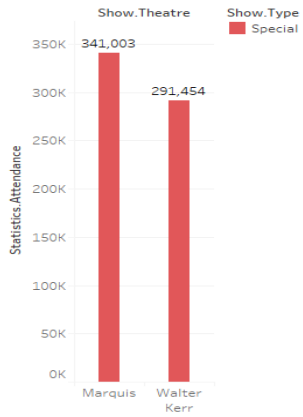
From the above charts, we see that there are more special performances in Walter Kerr and more capacity in comparison to Marquis. From the chart below, it is interesting to see that there is more attendance for special performances in Marquis than Walter Kerr. This trend is also seen in other cases. This indicates that few other external elements like location of theatre or quality of performances in a theatre could play a role in influencing the audience. The first visualization shows the data for all theatres and the second visualization corresponds to the specific example cited.

Theatres Vs. Attendance



Sum of Statistics.Attendance for each Show.Theatre. Color shows details about Show.Type.

Marquis/Walter Kerr Vs. Attendance



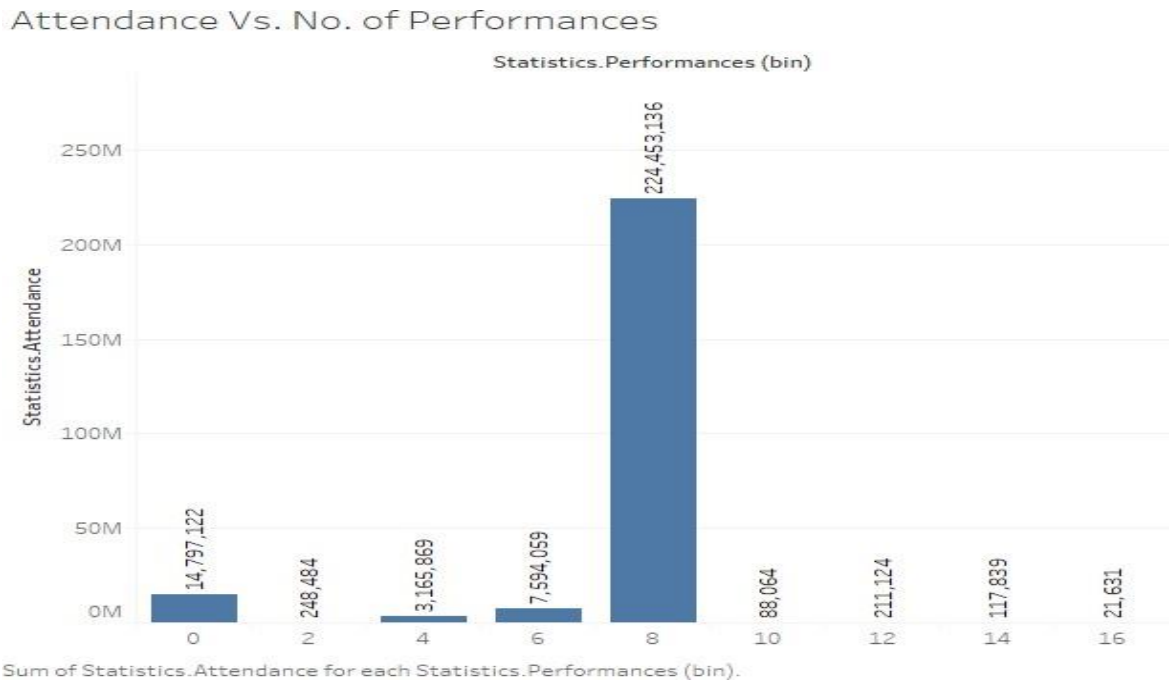
Sum of Statistics.Attendance for each Show.Theatre. Color shows details about Show.Type. The view is filtered on Show.Theatre and Show.Type. The Show.Theatre filter keeps Marquis and Walter Kerr. The Show.Type filter keeps Special.

It will be of interest to analyze the dataset in terms of logistics too.

How many performances should be held every week?

The number of performances held each week is binned by 2. The corresponding attendance of audience who attended the performances over the week is plotted. It is evident that when the number of performances per week is increased above 8, the audience attendance does not improve. Having lesser than 8 performances is also not advisable. It is best to have 8 performances each week to attract maximum attendance.

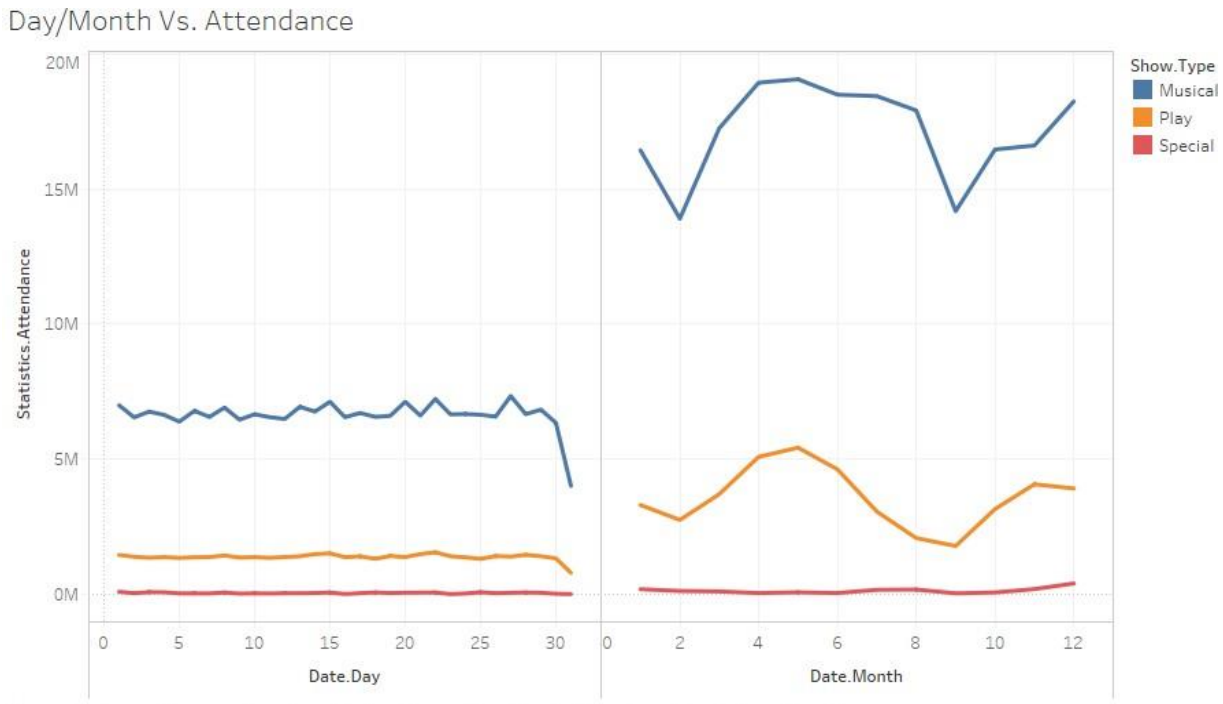
The attendance value shown for the number of performances in the 0 bin is relatively high. On closer inspection at the dataset, it exposes the flaw that few shows have non-zero attendance while having zero (no) performance in the week.



When is the off-season at Broadway?

When we look at the trend of attendance over the days and months, it is obvious that the variations within the days of the month are less prominent. The steep fall beyond 30th could be because 31st does not exist for all the months. Otherwise, across the different show types, there are low fluctuations in attending performances.

Across the months, there are two visible off-seasons in February (2) and September (9) for the musical and play performances. The special performances appear stagnant throughout the year, but it could be because of lower number of performances in general and that it is compared with a common scale. The minor trends could be overlooked. The fall in attendance in February could be attributed to it being the shortest month. While September could be seen as the beginning of fall and hence lower audience turn out. Similarly, the month of May (5) being the peak of Summer could have helped in higher audience attendance.



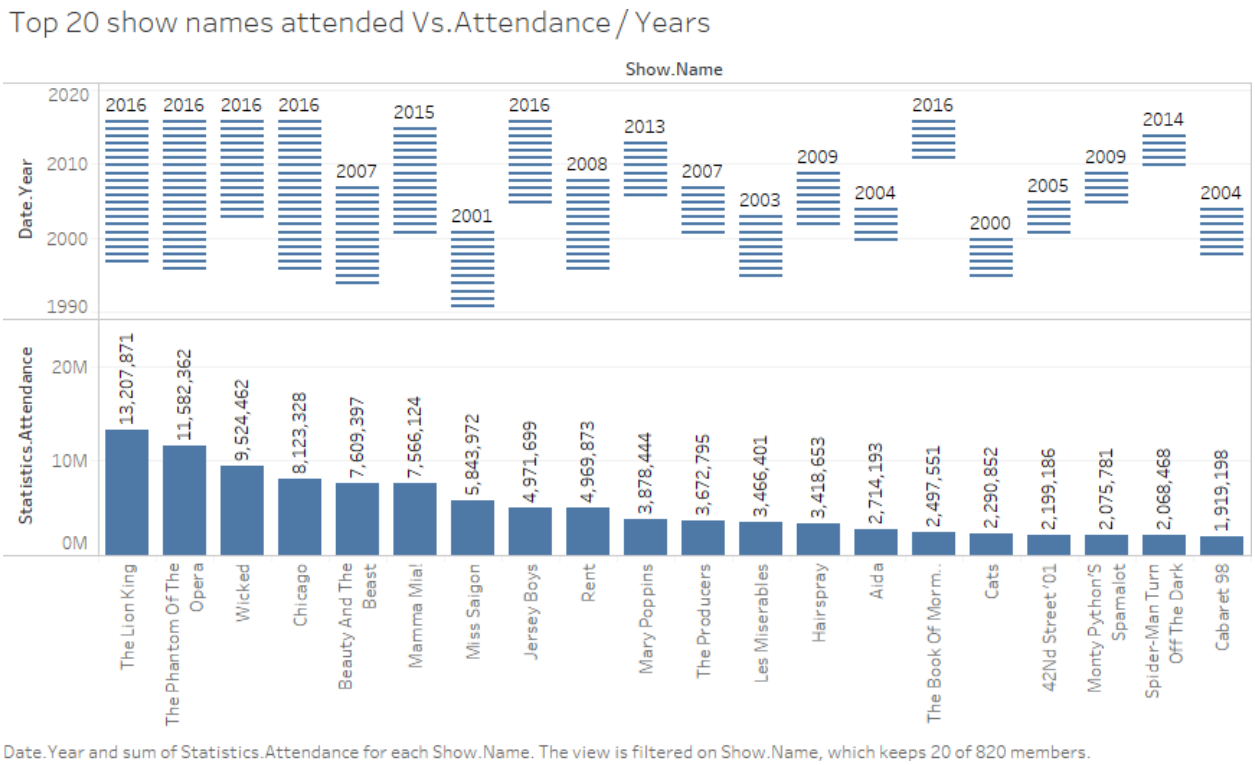
The trends of sum of Statistics.Attendance for Date.Day and Date.Month. Color shows details about Show.Type.

Which are the top shows that enjoyed the maximum audience attendance?

It is interesting to see what are the top 20 shows out of the whopping 820 shows that pulled in the maximum audience and which was the latest year they were performed.

In addition to that, from the Gantt chart for the years they were performed, we get an insight on how long they have been performed.

“The Lion King” tops the list followed by “The Phantom of the Opera”. It can be seen that “The Lion King” was performed an year later than “The Phantom of the Opera” but still managed to gather more audience.

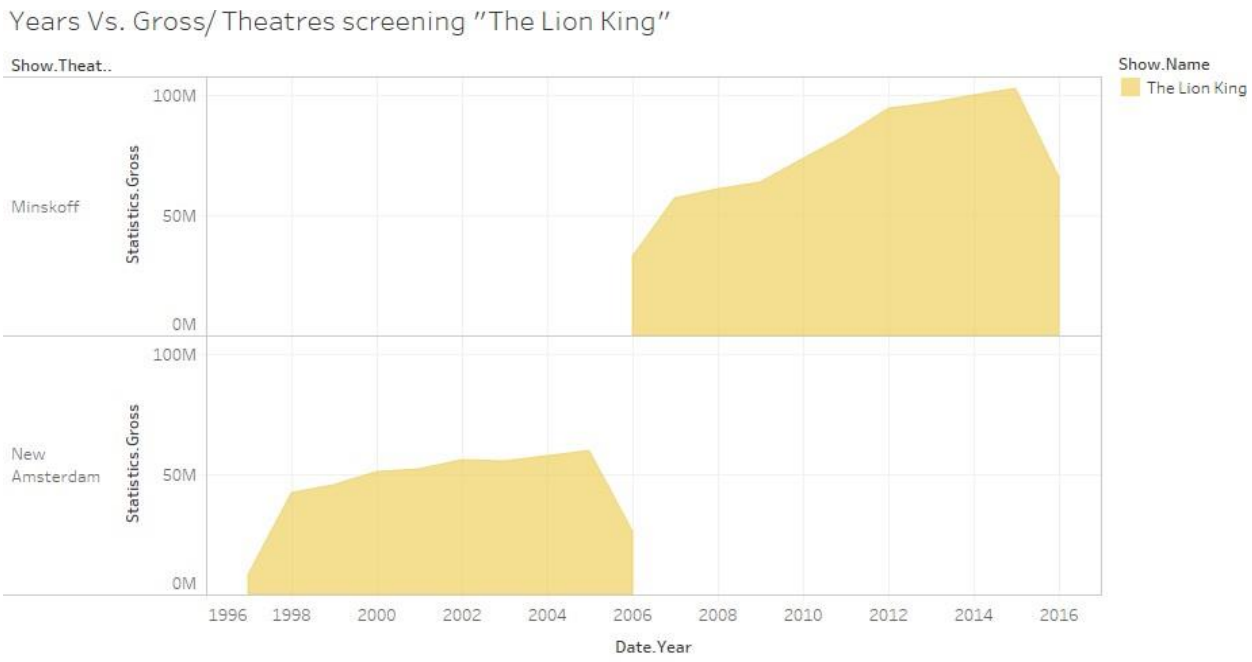


Date.Year and sum of Statistics.Attendance for each Show.Name. The view is filtered on Show.Name, which keeps 20 of 820 members.

Now that we know the people preferred watching “The Lion King”, it will also be of interest to know which theatres catered to the screening of this popular show.

Where did people see “The Lion King”?

From this visualization, it is evident that only two theatres screened the most popular “The Lion King”. Initially, it had been performed at New Amsterdam from 1997 to 2006 and later was performed at Minskoff until 2016. The Gross yield as expected had increased over time reaching a peak at 2015. The drop observed in 2016 could be due to lack of complete data. The drop at 2006 levels in New Amsterdam is because, it was being performed at both the theatres in 2006.



The plot of sum of Statistics.Gross for Date.Year broken down by Show.Theatre. Color shows details about Show.Name. The view is filtered on Show.Name, which keeps The Lion King.

Summary:

From the dataset, exploratory data analysis was done, and few important and interesting insights were gained.

1. Musical performances made the highest gross yield.
2. 2014 was the peak of gross yield.
3. Data corresponding to 2016 might be incomplete.
4. Gross potential has been an effective prediction reflecting audience attendance since 1998.
5. I am more likely to get a ticket at Gershwin followed by Majestic owing to its capacity.
6. Probability of watching a musical performance is highest at Majestic. For plays, it is Ethel Barrymore and for special performances, it is Walter Kerr.
7. More number of performances and capacity alone does not guarantee attendance.
8. It is best to have 8 performances each week to attract maximum attendance.
9. February and September are off seasons at Broadway. May attracts maximum crowd.
10. "The Lion King" tops the list of 820 shows which got the maximum attendance. It was last performed in 2016.
11. "The Lion King" was performed only at two theatres. It yielded maximum gross at Minskoff where it was performed post 2006.

This analysis makes understanding large amounts of data stretching over 50+ theatres with 820 shows spread between 1990 and 2006 through each week feasible.

It also indicates potential flaws in data collection like:

1. Incomplete data for the year 2006,
2. Zero gross potential predicted for selected years and
3. Zero performances for shows with non-zero audience attendance.

Hence, this exploratory Data Analysis process has been very effective in gaining useful insights from the Broadway Dataset.