Report: ANLY 501 Project Part 2

Lujia Deng, Luwei Lei, Janet Liu, Yunfei Zhang
ld781,ll1038,yl879,yz678@georgetown.edu

Monday, November 4, 2019

1 Exploratory Analysis

1.1 Basic Statistical Analysis & Data Cleaning Insight

1.1.1 Basic Description: Mean, Median, and Mode

The mean, median, and standard deviation of selected attributes from the *Broadway Grosses* and *Broadway Social Media Stats* data sets are shown below:

attributes	*mean*	*median*	*standard deviation*
Likes Vs.Last Week	798.70	150.00	4,605.80
cleaned_FB Talking About	7,643.38	3,770.00	18,890.98
Checkins vs.Last Week	472.73	442.00	1,496.07
Twitter vs.Last Week	297.24	76.00	1,463.33
IG Followers vs.Last Week	1,219.06	504.00	2,560.83
Weekly Grosses	568,324.59	467,067.00	435,518.20
Average Ticket Price	67.29	59.81	37.87
Seats Sold	7,898.15	7,732.00	3,168.94
Percent of Cap	0.80	0.83	0.17
Number of Performances	7.24	8.00	2.23

Table 1: The mean, median, and standard deviation of attributes in the two data sets.

As can be seen in Table 1, the mean of Like v.s. last week is 798.70 and the median is 150.0, which indicate that the mean of Facebook likes of each show increased 799 in average. However, it is likely that there are some outliers in the data set as the median was 150, which is far less than the mean. The same pattern has been found in the attributes Facebook Talking about and Twitter/IG Followers v.s. last week as well. On average, all the shows had 7,643 talking about on Facebook each week, with a mean of 3,770, and a relatively high standard deviation of

18,891. The increase of followers in Twitter was 297, with a median of 76 and a standard deviation of 1463; the increase of followers on Instagram was on average 1219, with a median of 504 and also a high standard deviation. By comparing the Facebook likes, Twitter and IG followers, we could intuitively infer that Broadway shows draw more attention on IG than on Twitter and Facebook.

In terms of attributes related to the sales, most of them have similar means and medians, but some have large standard deviations compared to the means, which indicate large variances in sales data of different shows and time. For example, the average of weekly grosses is $568,324 and the median is $467,067, but the standard deviation reaches $435,518, which is close to the mean. It implies extreme values that significantly deviate from the mean and a possible skewed distribution. Similar patterns can be observed in the average ticket price, whose mean, median, and standard deviation are $67, $59 and $37, as well as in the number of seats sold, with the mean, median and standard deviation as $7,898, $7,732, and $3,168. The average percent of cap is 80% and the average number of performance per week is 7. Both attributes have relatively small standard deviations compared to the means.

1.1.2 Data Cleaning

1. Outliers

   To identify outliers, we took a combined approach with both visualizations and the Local Outlier Factor algorithm (LOF), an unsupervised anomaly detection algorithm. Since the grosses data set is a stack of time series data, it is intuitive for us to explore potential anomalies over time. We picked two long-lasting Broadway shows, *Wicked* and *Mamma Mia!*, and plotted their weekly grosses and average ticket prices over time.

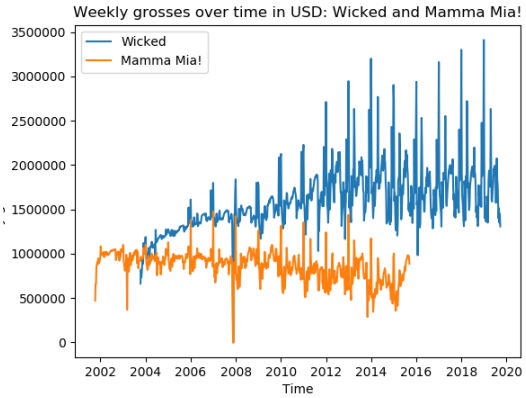

Figure 1: Weekly Grosses Over Time of *Wicked* and *Mamma Mia!*

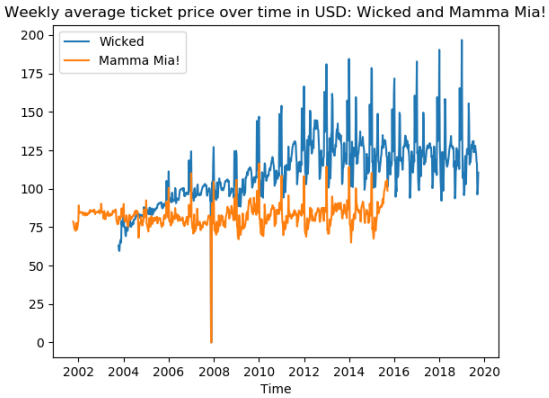

Figure 2: Weekly Average Ticket Price Over Time of *Wicked* and *Mamma Mia!*

In addition to the seasonality effects observed from the plots, it is clear that there is a significant drop in the price and gross of both shows at the end of 2017. As we delved deeper into the historical background, we discovered that a Broadway stagehand strike had took place in

the week of 11/25/2017, where all performances were cancelled and 22 shows were affected, including *Wicked* and *Mamma Mia!*, exactly corresponding to our data. We decided to keep these interesting outliers and marked them out with the value of -1 in the column called 'strike'.

For both the *Broadway Grosses* and *Broadway Social Media Stats* data sets, We applied LOF and detected outliers of each show. As we set the numbers of neighbors as 5, the shows that have less than 5 records would not be included in the outlier detection. Around 10% of the records were detected as outliers and labeled as "-1" in the data set.

Outliers identified in the *Broadway Grosses* data set are not very informative. We tried to plot them over time (shown in Figure 3), but anomaly detection by LOF in the high dimensional space is hard to be observed in the 2D space.

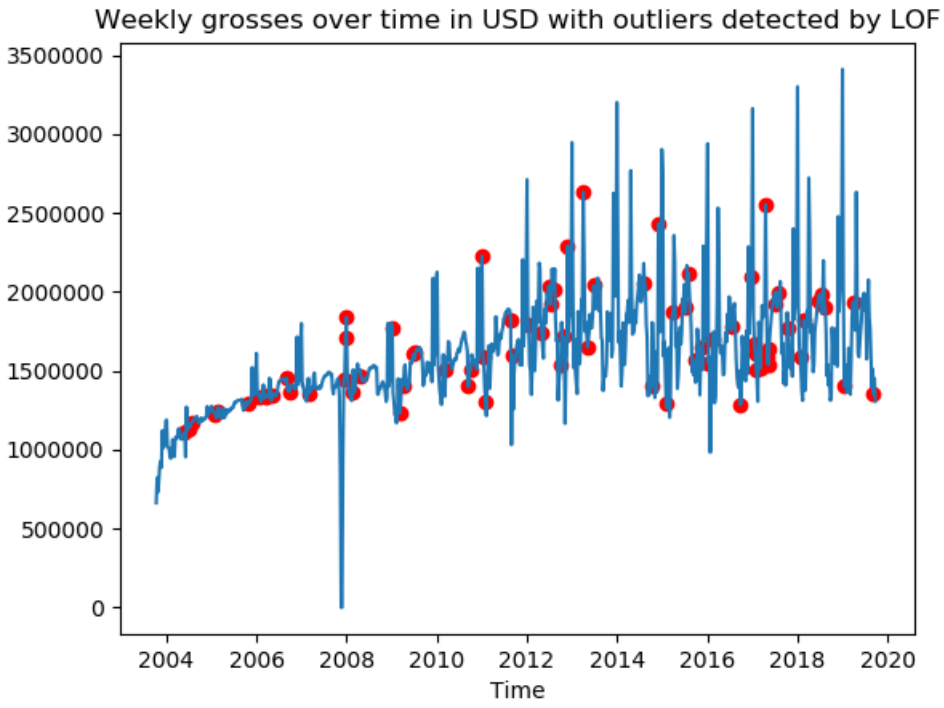

Figure 3: Weekly Grosses Over Time of *Wicked* with Outliers

For the *Broadway Social Media Stats* data set, we have found the two following patterns:

- It might be the first week of a show to be recorded in the data set, and thus the value comparing to last week (e.g. CHECKINS VS. LAST WEEK) would be the exact the same as CHEKINS THIS WEEK. This should not be seen as an error, but this indeed has some different patterns from other records and should not be removed.
- For some records, the shows received more attention in a week, which might be due to the

marketing activities or news. This pattern would be interesting to be further explored and should not be deleted.

It is worth noting that there are seemingly odd data that can be justified in the context of Broadway. PERCENT OF CAP (i.e. percentage of seats sold out of total seats) greater than 100% may seem suspicious, but it indicates over-sales, a common business strategy in the Broadway business. Top shows can be extremely popular and generate disproportionately large grosses compared to others. Additionally, seasonality is also a significant determinant of the sales. For instance, the maximum weekly grosses data point, $4,041,493, belongs to the history maker Broadway show *Hamilton* in the holiday season 2018 (the week of 12/30/2018). Because these 'extreme' values reflect the nature of the Broadway business, we decided to keep them in our data set.

2. Missing Values

   In this part of data cleaning, missing values are also handled. For the *Broadway Social Media Stats* data set, the initial data set contained some errors and duplication, which were solved in Project Part 1. This data set contains some "0" values, and some of them should be treated as missing values, while others are just exact zeros. For example, the attribute FACEBOOK LIKES is accumulative, which is expected to be increasing over time. Hence, in this case, "0"s were not expected. FB CHECKINS had this similar issue, so we reassigned the mean value of the show to the missing values. However, it should be noted that, when a show had not come out, it might have FACEBOOK TALKING ABOUT, but not CHECKINS. As a result, we only reassigned the missing values for shows that were currently available.

   For the *Broadway Grosses* data set, we handled the missing values in Project Part 1. We have reiterated a few points here. There are two cases of missing values in this data set. One is when a record consists of null values, and the other one is when a row is full of zero values. Previously we mentioned that some of zero values are due to holiday seasons, but at this stage, we found out that those are actually results of a stagehand strike in 2007, so we make a correction here. As mentioned above, we decided to keep these data in our data set with a marker in the 'strike' column. Other than that, zero values are indications of missing data. Therefore, we removed null values and records of shows with only zero grosses and statistics, which was done in Project Part 1.

3. Other Data Cleaning Decisions

   We recalculated some attributes in the *Broadway Social Media Stats* data set. There are some attributes that describe the comparison versus last week. For example, we have both IG FOLLOWERS and IG FOLLOWERS V.S. LAST WEEK. The comparisons with last week provided us with a lot of insights, especially when connecting them with gross change and relative news. As we cleaned our data set, we recalculated these attributes to make sure they are correct. We also converted the dates in the grosses data set from strings to data/time objects and generated

new features as year and month, in order to incorporate time effects into our anomaly detection mentioned above and hypothesis testing model, which will be discussed in Section 2.1.

1.1.3 Smoothing Data Using Binning

We have chosen to bin the following two attributes:

1. THIS_WEEK_GROSS

   We would like to evaluate a show's ability to generate revenues, and weekly gross is the most important indicator of sales in our data set. Instead of paying attention to the exact number of grosses, we are more interested in the order of magnitude. The question to ask is: is the show a $100K-level or a $1M-level one? Therefore, it makes sense to us to bin the attribute THIS_WEEK_GROSS.

   The bins are chosen based on the order of magnitude as well as the distribution of the data. Most of the data are above 100,000, so the preliminary bins are <100,000, 100,000-1,000,000, and >1,000,000. Since there are few data with values of zeros (due to the strike), we isolated them by binning them together, using a bin from (-1, 1]. After trying out this strategy, we noticed that there were too many observations gathering in the bin of (100,000, 1,000,000], so we decided to cut it in half; as a result, the resulting bins and number of data points associated with each are shown as below:

bins	*number of data points*
(-1, 1]	53
(1, 100000]	2,648
(100000, 500000]	22,510
(500000, 1000000]	15,355
(1000000, 10000000]	6,159

Table 2: Binning Result of the Attribute THIS_WEEK_GROSS.

2. PERCENT_OF_CAP

   The percent of cap measures the number of seats sold over total number of seats available per show. It indicates the popularity of a show, which can be useful in our analysis. Again, it is the general ranges rather than the exact percentages we are interested in, and the binning strategy helps us smooth out the noises. Since most of the shows have a percent of cap over 50%, and a value below this threshold may be a strong indicator of unpopularity, we binned observations below 50% together. For data between 50%-100%, We set the bin width as 10%, resulting 5 bins. The two special bins, (-1, 0] and (1, 2] are also designed to account for cases of 0 values and over-sales. The final bins and number of data points associated with each are shown as below:

bins	*number of data points*
(1.0, 2.0]	4,416
(0.9, 1.0]	12,443
(0.8, 0.9]	9,347
(0.7, 0.8]	8,153
(0.6, 0.7]	6,000
(0.5, 0.6]	3,601
(0.0, 0.5]	2,709
(-1.0, 0.0]	56

Table 3: Binning Result of the Attribute PERCENTAGE_OF_CAP

1.2 Histograms & Correlations

From the histograms below, we can observe that three important quantitative attributes in our data set are not normally distributed. WEEKLY GROSSES and AVERAGE TICKET PRICE are right-skewed and PERCENT OF CAP is left skewed.

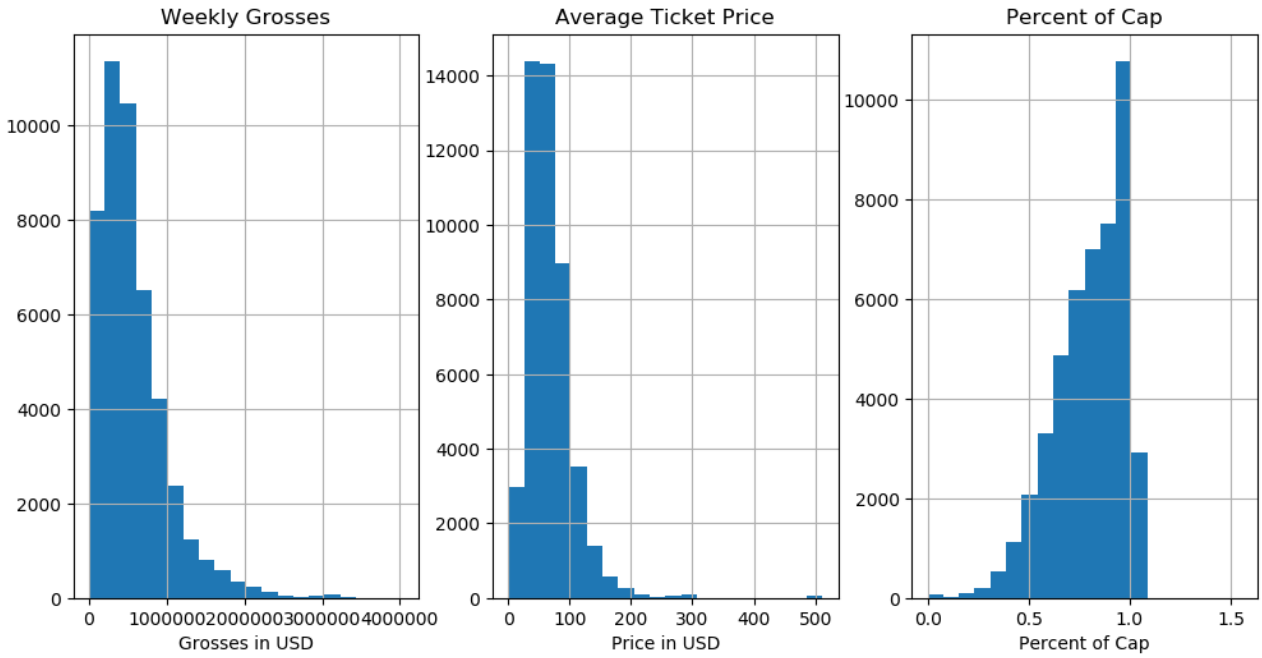

Figure 4: Histograms of WEEKLY GROSSES, AVERAGE TICKET PRICE, and PERCENT OF CAP

The data correspond with our prior knowledge of Broadway shows. Most show generate revenues between $100,000 and $1,000,000, and there are also extremely popular productions that can make disproportionately high revenues, which are shown by the long tail from the first plot in Figure 4. Similarly, AVERAGE TICKET PRICES are centered around $50, but we can also see a considerable

amount of shows sold at a price range between $100 and $200. For popular shows such as *Hamilton*, the price can reach $300 - $500.

From the histogram of PERCENT OF CAP in Figure 4, we can observe that it is relatively rare for a show to have a percent below 50%. One possible reason is that theater managers would cut off a show before its percent of cap drops below 50%. Most shows have over 70% percent of cap, and some of them can be oversold, resulting in a percent above 100%.

We are interested in the relationships between seats sold, average ticket price, and percentage of capability (i.e. tickets sold / the total capability of the theatre). Were cheaper shows more popular? Or did expensive shows indicate a blockbuster so they were better sold? The correlation table and plots would provide interesting insights.

variable	*seats_sold*	*avg_ticket_price*	*percent_of_cap*
SEATS_SOLD	-	0.31	0.63
AVG_TICKET_PRICE	-	-	0.43
PERCENT_OF_CAP	-	-	-

Table 4: The Correlation Table of Variables SEATS_SOLD, AVG_TICKET_PRICE, and PERCENT_OF_CAP.

As shown in Table 4, the correlation between seats sold and average ticket price is 0.37, which indicates a mild positive relationship between these two variables. In general, popular shows have higher ticket price. This may due to the fact that the price itself was higher, or much higher price tickets were sold in these shows. The similar result can be found between average ticket price and percentage of capability, where the correlation score is 0.43. Seats sold and percentage of capability are highly correlated, with a r equal to 0.63, because both of them can be seen as measurement of popularity of the show but in different dimensions.

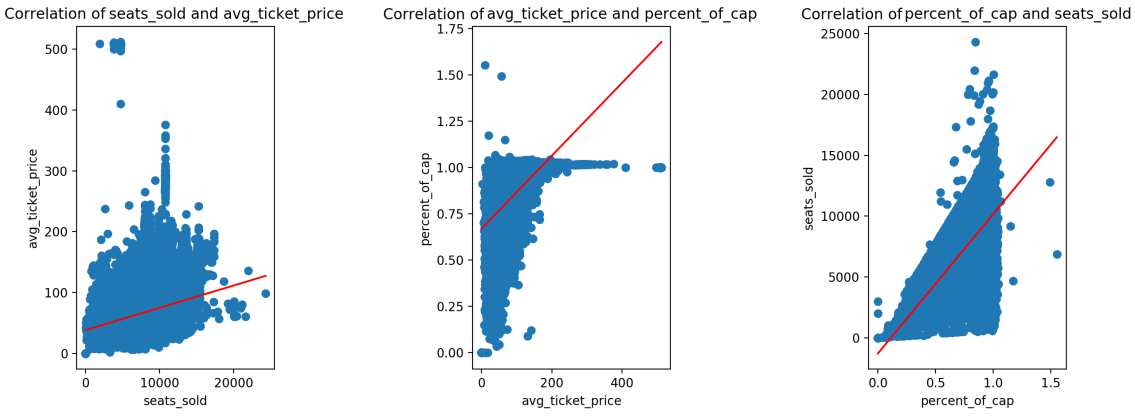

Figure 5: Scatter Plots of Seats Sold, Average Ticket Price and Percent of Cap.

We also plotted the relationships of the aforementioned 3 variables in a set of subplots, as shown in Figure 5. The red line is the best fit line of the correlationship. As can be seen in the plots, though

the data points are not closely distributed beside the best fit line, positive correlation can be found in those variables.

1.3 Clustering Analysis

We conducted three three cluster analyses on our *Broadway Social Stats* data set. In particular, we looked at 4 attributes, which are shown in Table 5. The three clustering methods are WARD, K-MEANS, and DBSCAN, and we applied both the Silhouette and Calinski-Harabaz procedures to assess the quality of the clusters. The results are shown in Table 6. In addition, Figure 6, Figure 7, and Figure 8 show the scatter plot of the clusters using each clustering method.

For this clustering analysis, we looked at the fluctuation in LIKES and FOLLOWERS on three different social media platforms (Facebook, Twitter, and Instagram) in relation to whether a show is current on air or upcoming. As can be seen from the figures, the three clustering methods produced similar clustering shapes. In particular, the WARD clustering method is a agglomerative hierarchical clustering method that does not have to assume any particular number of clusters. However, once a decision is made to combine two clusters, it cannot be undone. Moreover, no global objective function is directly minimized. As a result, it produced the lowest *Silhouette* score, indicating a low quality of the clusters. According to the quality scores in Table 6, the K-MEANS clustering method produced the best performance. The scores resulted from determining the number of cluster as 3 (i.e. k=3) after experimenting with different cluster numbers. It is true that K-Means will lead to the lowest quality clusters if there is noise in the data set. However, the data set being clustered here is clean, and that's why K-Means performed pretty well. DBSCAN is a density-based algorithm and can handle noise best. Thus, its performance on the *Silhouette* procedure is between the other two clustering methods. However, its performance on the *Calinski-Harabaz* procedure is the worst. DBSCAN can identify points that are not part of any cluster (very useful as outliers detector), which are the dark blue points in Figure 8. Similar to WARD and unlike K-MEANS, DBSCAN does not require us to set the number of clusters *a priori*. Furthermore, the plot of DBSCAN seemed to provide a clear view of how the data points are clustered, despite its low score from the *Calinski-Harabaz* procedure.

In addition, as can be seen from Figure 7 and Figure 8, the clustering method K-MEANS identified one cluster on the right (i.e. shown in yellow) whereas DBSCAN identified such as two clusters (i.e. shown in orange and purple), which is likely due to the fact that these data points are the most popular/unpopular shows compared to the previous week, thereby resulting in the gap.

attributes	*type*	*values*
CURRENT	categorical	*Current* or *Upcoming*
LIKES VS.LAST WEEK	numerical	numerical values
TWITTER VS.LAST WEEK	numerical	numerical values
IG FOLLOWERS VS.LAST WEEK	numerical	numerical values

Table 5: Attributes used in the *Broadway Social Stats* data set for Clustering Analysis.

	Silhouette	*Calinski-Harabaz*
WARD	0.73	726.83
K-MEANS	0.91	1188.03
DBSCAN	0.85	326.81

Table 6: The Quality of the Clusters using both the Silhouette and Calinski-Harabaz Procedures.

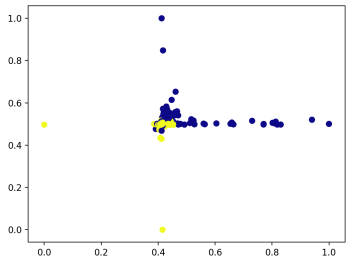

Figure 6: Scatter Plot for the Clustering Method WARD.

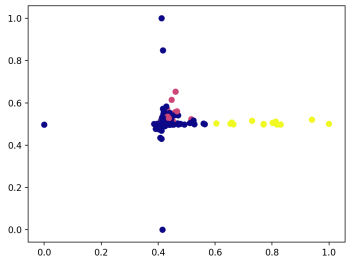

Figure 7: Scatter Plot for the Clustering Method K-MEANS.

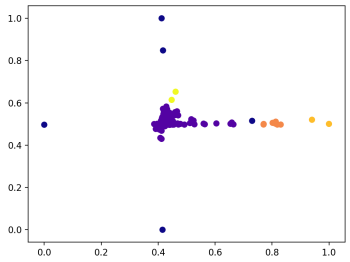

Figure 8: Scatter Plot for the Clustering Method DBSCAN.

1.4 Association Rules / Frequent Itemset Mining Analysis

After studying the distributions of the *Broadway Social Stats* data set, we would like to take a look at the associations among popular shows. In particular, we want to figure out which shows have higher sales at the same time period. This could provide us some potential insights into customer tastes.

The transaction data are defined by collecting the show names of each week that have at least 80 percentages of seats sold out of total seats. We avoid to condition on actual monetary sales since the inflation rate would affect the sales in dollars. After generating 1,790 records of transaction and applying Apriori algorithm with support values of 0.4, 0.6, and 0.8, we only found out some monotonic associations between shows. The itemsets with support and confidence values are saved in ITEMSET_SUPPORT0.CSV and ITEMSE_CONFIDENCE0.CSV respectively. We noticed that the generated itemsets mostly only contain the famous and long-lasting shows. For example, *The Phantom of the Opera* runs from 1988 to present, which is the longest running show in Broadway history. It outputs the highest support value, 0.923463687150838. Other shows such as *Mamma Mia!*, *Chicago*, *Les Misérables*, *The Lion King* [lio] have at least ten-year performances on Broadway. We then observed the confidence values among association rules. Most of the rules have a very high confidence. For instance, the level of confidence of *The Lion King* and *Mamma Mia!* conditioning on *The Phantom of the Opera* has 1.0 confidence. We double checked the years of productions of these shows and found out that since all of them are the top longest-running show on Broadway, their time ranges are overlapping. Therefore, the confidence values are very high.

In order to observe some more general associations between newer performances of shows, we

modified the time range from 1985-2019 to 2000-2019. This produced many more itemsets than the previous case. The results are saved in ITEMSET_SUPPORT1.CSV and ITEMSE_CONFIDENCE1.CSV respectively. The association rules produced some interesting insights this time. While the long-running shows still achieved very high support values, the newer shows such as *Jersey Boys*, *Rent* and *The Book of Mormon* started demonstrating their significance. We researched that these shows also have been running several years, which means that the popular and classic shows would attract more audience [sho]. The largest support value is 0.9979529170931423, given by rule ('The Phantom of the Opera','Chicago'). These two are the top two longest-running Broadway shows, which is not surprising that they are mostly chosen by audience. The confidence values of most association rules are larger than 0.8. The famous and classic shows are always welcomed by audience during the same time period.

2 Predictive Analysis - Part 1

2.1 Hypothesis Testing

Hypothesis 1 Seats sold during Thanksgiving week, winter holiday week, and summer break months spike as opposed to the rest of the time during the year.

Method one-way ANOVA analysis

Data Preparation Categorize date into different time periods

Description ANOVA is used to compare the difference in mean among more than two groups, which is the case of this hypothesis. The seats-sold records are aggregated by week in our original data set. To do the ANOVA analysis, we first labeled them to 4 categories using the date (i.e. the attribute WEEK_ENDING). We designed a simple yet smart algorithm to find the holidays. Taking Thanksgiving as an example, this holiday is on the last Thursday of each November, so the possible date would be 24th - 30th November; thus, the end of Thanksgiving week would be Nov.27 - Dec.3, which were used to locate the Thanksgiving week in our data set. Besides, as people start to travel the week before Thanksgiving, we labeled that week as "Thanksgiving" for the analysis, too. We applied this method to label the records into 4 categories: Thanksgiving, WinterBreak, SummerBreak, and NotHoliday.

Data Analysis ANOVA and Post-hoc Pairwise Analysis
We applied ANOVA analysis to detect the difference in means among the seats sold of 4 time periods, and obtained a p value of 0.016, which indicates that there is a difference in the mean amount of seats sold among the 4 groups. Detailed analysis is shown in the script.

As we obtain the result that the seats sold among time periods are significantly different, a post-hoc analysis was applied to detect which 2 groups have differences in mean.

time periods	*avg weekly seats sold*
NOT HOLIDAY	207,240.33
SUMMERBREAK	199,135.31
THANKSGIVING	224,891.53
WINTERBREAK	206,807.24

Table 7: The summary of average weekly seats sold of different time periods.

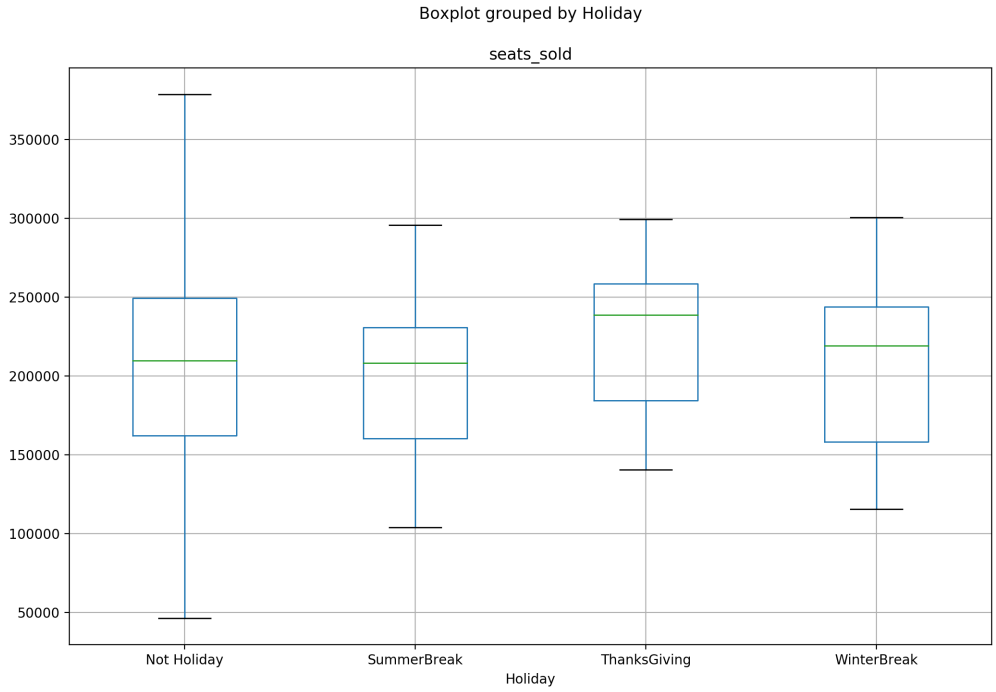

Figure 9: Boxplots of Seats Sold by Holiday

Results & Analysis As can be seen in the table above, the seats sold over summer breaks are the lowest and those of Thanksgivings are the highest. According to the pairwise analysis, the only significant result is the difference between summer break and thanksgiving.
Our initial hypothesis was partially supported. Thanksgiving is indeed a profitable season for Broadway. But conversely, summer breaks sold less tickets, which was not as expected. Thus, summer time might be interpreted as a low season of Broadway shows.

2 groups	*meandiff*	*p*	*lower*	*upper*	*reject*
Not Holiday v.s. SummerBreak	-8,105.01	0.07	-16,635.54	425.51	False
Not Holiday v.s. ThanksGiving	17,651.20	0.22	-5,922.80	41,225.21	False
Not Holiday v.s. WinterBreak	-433.091	0.90	-24,007.09	23,140.91	False
SummerBreak v.s. ThanksGiving	25,756.22	0.04	1,212.63	50,299.80	True
SummerBreak v.s. WinterBreak	7,671.92	0.83	-16,871.66	32,215.50	False
ThanksGiving v.s. WinterBreak	-18,084.29	0.49	-51,028.96	14,860.37	False

Table 8: The Pairwise Analysis to Compare Pairs of Two Groups.

Hypothesis 2 If a show is a money maker, it is highly possible that it also has a high general rating (a combination of critics and readers' ratings).

Method Logistic Regression

Data Preparation & Description To test this hypothesis, we merged grosses and ratings data. We also constrained the time scope to be the recent 5 years in order to limited the time effects on grosses (grosses have been significantly increasing over years). A high rating is defined as a score greater than 7. For each show, the gross data are aggregated by taking the average value of all the data points and normalized using a min-max scaling approach. The hypothesis testing was performed using a logistic regression model because the dependent variable is binary. The conceptual model is shown as the following: ratingLevel(0, 1) = $X_0 + \beta_1 *$norm_gross.

Results & Analysis The resulting coefficient and p-value for norm_gross is 2.98 and 0.009. In general, we can see that there is a positive correlation between grossed and ratings. The p-value shows that our result is statistically significant. Through exponentiating the coefficient, we got the odd-ratio 19.66, meaning that for one unit of increase in the norm_gross, the odd of the show being highly rated increases by a factor of 19.66.

We have also realized that only incorporating one factor into the model leaves other factors uncontrolled, therefore weakening the explanatory power of our model.

Hypothesis 3 The larger a theater is, the higher the average ticket price of the show will be.

Method Linear Regression

Description Large theaters tend to have better infrastructure and host large musical productions, which are usually long-running and popular. *Wicked* and its home the Gershwin Theater and *Phantom of the Opera* and its home the Majestic Theatre are all good examples. Due to larger productions and better infrastructure, it is possible that the average prices for shows played in large theaters are

higher. However, there are also many legendary small theaters, such as the Helen Hayes Theater and the Circle in the Square Theatre, which housed many notable Broadway shows. It is also fair to assume that the average price would be higher due their high reputation and limited capacity. The combined positive and negative effects of the size of theaters on ticket prices piqued our interest, and we hoped a hypothesis testing can give us a preliminary answer.

Data Preparation The size of the theaters is calculated using the following formula:

Number of seats = (weekly seats sold / percent of cap) / number of performances per week

We also conducted a log transformation of the number of seats and average ticket price to adjust their skewed distributions in order to fit them into a linear regression model. We also incorporated year and month attributes into the model, because we want to control for the season effect on the ticket price. The conceptual model is as the following:

$\text{log\_average\_ticket\_price} = X_0 + \beta_1 * \text{log\_num\_seats} + \beta_2 * \text{month} + \beta_3 * \text{year}$

Results & Analysis The result shows that a coefficient of 0.1805, with a p-value close to zero, is associated with log_num_seats, indicating a positive relationship between the theater size and the average ticket price. Our R-squared scores 0.68, implying that about 68% of the variance of the data can be explained through our model.

2.2 Classification

2.2.1 Classification Task 1: Grosses Prediction using DT, KNN, SVM, and RF

This classification task is focused on using the social media data in previous week to predict if gross increased or not in this week. We believe that the social media reactions of each show could somehow impact its gross in that there is a possibility that the audience's willingness to attend the show may be affected by its reviews and popularity on social media.

We created the labels of gross by converting the gross differences comparing to previous week to binary attribute. If the difference is positive, it is assigned as 1, otherwise, as 0. Then, we matched the dates by subtracting seven days from social media dates and joined with the gross difference attribute. Since the data set contains large number of features and it's relatively small, we decided to try Decision Trees, Random Forest, SVM and KNN methods for this task.

The training and testing accuracy for each method is recorded in Table 9. Each method except Random Forest performed a ten-fold cross validation to better predict the gross. Random forest was set to one hundred estimator, which is the number of trees in the forest. It achieved the highest training score; however, this might be an over-fitting case since the testing accuracy is only approximately 0.6. The KNN model had the highest testing accuracy among four different models. We then took a look at the confusion matrix of each model on the testing data. The decision tree and random forest have similar distributions of prediction results, while Support Vector Machine

	DT	KNN	SVM	RF
Training Accuracy	0.565383	0.542322	0.566995	0.858001
Testing Accuracy	0.592105	0.605263	0.546052	0.598684

Table 9: Model Performances on the Grosses Prediction Classification Task.

	DT		KNN	
	Predicted 0	Predicted 1	Predicted 0	Predicted 1
Actual 0	44	31	51	24
Actual 1	36	41	36	41
	SVM		RF	
	Predicted 0	Predicted 1	Predicted 0	Predicted 1
Actual 0	64	11	45	30
Actual 1	58	19	31	46

Table 10: Results of Each Label from Different Classification Models.

have more biased results. The K-nearest neighbor has the highest precision in this classification task comparing to others.

We plotted the ROC curve for the decision tree model, as shown in Figure 10. Similar to the confusion matrix results, the decision tree model hardly contributes to predicting the gross. Its true positive rate and false positive rate are positively correlated. We also produced a feature importance graph from random forest model, as shown in Figure 11. The attribute DATE plays the most important role in prediction, followed by several social media reactions comparing to last week. This suggests that social media reactions of previous week may affect the attendance of the show, but they are not the key factors since the overall accuracy is low.

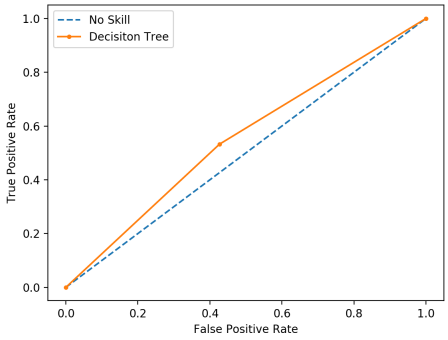

Figure 10: The ROC Curve for the Decision Tree Model.

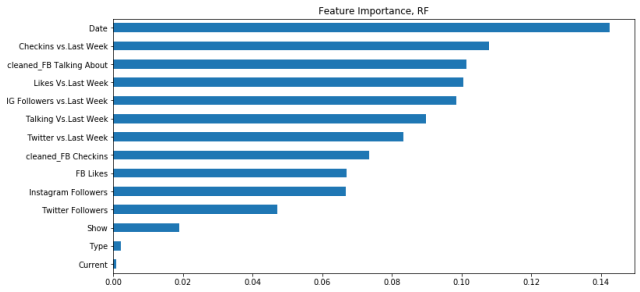

Figure 11: Feature Importance Graph from the Random Forest Model.

Even though we tried to increase the performance by modifying parameters and applying cross

OVERSOLD	POPULAR	UNPOPULAR
(1.0, 2.0]	(0.5, 0.6], (0.6, 0.7], (0.7, 0.8], (0.8, 0.9], (1.0, 2.0], (0.9, 1.0]	(-1.0, 0.0] (0.0, 0.5]

Table 11: A Detailed Description of the Labels and their Corresponding Binning Results.

validation, all of these results are not very promising. We concluded that the social media reactions do not have a direct impact on grosses of the following week.

2.2.2 Classification Task 2: Popularity Classification using Naive Bayes

For this classification task, we used the *Broadway Grosses* data set to predict a show's popularity given its number of performance in a week (i.e. the attribute PERFS). Specifically, we have created three labels to define such popularity: OVERSOLD, POPULAR, and UNPOPULAR. These labels are categorized and assigned based on the the binning results on the attribute PERCENT_OF_CAP, meaning the percentage of seats sold out of the total seats, which are shown in Table 11. The results are shown in Table 12.

As can be seen from Table 12, the label OVERSOLD has the lowest f1-score (0.37) and the label POPULAR has the highest f1-score (0.82). These results seem to reflect the distribution of each label in the data set: there are 39,544 records of the label POPULAR, 4,416 records of the label OVERSOLD, and 2,765 records of the label UNPOPULAR. In addition, the label OVERSOLD has only one corresponding binning categories whereas the label POPULAR has six corresponding binning categories, as shown in Table 11. The accuracy score of the testing data suggested that the number of performance in a given week is relatively indicative of the popularity of a show to some degree. As a result, when considering the popularity of a show, the feature PERFS can help us obtain some insights. However, since the performance is not very well, it would be interesting to add more features to this classification task in order to see performance fluctuation across the board.

3 Interim Conclusion

As we described in Project Part 1, the data science problem we are interested in is how the reputation of a Broadway show is represented in relation to its associated factors and what insights we could gain in order to benefit Broadway in a variety of ways.

Having conducted several exploratory and predictive analyses in this part of the project, we identified several factors that may or may not be indicative of a show's popularity and grosses such as *number of performance in a week, social media reactions, theater size, average ticket price* etc. We also identified the impact that seasonality has had on Broadway shows. This important factor cannot be directly observed from the data sets we collected. However, with our hypothesis testing as well as classification task, we have obtained some insights that explain our intuition. In addition, when we observed some abnormal fluctuations in the data sets as well as our analysis, we did some research

	Precision	Recall	f1-score
OVERSOLD	0.35	0.39	0.37
POPULAR	0.91	0.76	0.82
UNPOPULAR	0.27	0.86	0.41
accuracy			0.73
macro avg	0.51	0.67	0.53
weighted avg	0.82	0.73	0.76
Accuracy score of **training** data: 0.7237			
Accuracy score of **testing** data: 0.7274			

Table 12: Naive Bayes Performance on the Popularity Classification task.

on the historical background for the time period in question. For instance, the strike that happened in 2007 explained the irregularity demonstrated in our data sets and analyses. We aim to conduct more predictive analyses later in the project to further substantiate our findings and provide more valuable interpretations of the results.

References

[lio] The Lion King. `https://www.ibdb.com/broadway-production/the-lion-king-4761`. [Online; accessed 02-Nov-2019].

[sho] List of the longest-running Broadway shows. `https://en.wikipedia.org/wiki/List_of_the_longest-running_Broadway_shows`. [Online; accessed 02-Nov-2019].