Report: ANLY 501 Project Part 1

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

Friday, October 4 2019

1 Data Science Problem

Nowadays, data science has been employed in media and entertainment with the advancement of technology. Data analysis, especially within the entertainment industry, can be exciting work as it can shed light on what factors are affecting the prosperity of media and entertainment. For instance, Farsite Group, a data science firm based in Columbus, Ohio, launched a highly visible campaign in early 2013 to use predictive analytics to forecast the winners of the 85th Annual Academy Awards [2], which illustrated how data science could be further deployed in the media and entertainment industries.

Therefore, we would like to explore another representative industry along this line of work – Broadway, which is widely considered to represent the highest level of commercial theatre in the English-speaking world. Moreover, the business landscape of Broadway has become highly modernized with digital marketing, social media, e-ticketing, and dynamic pricing at the cutting edge of technological advancements within entertainment [1], which provides various types of data and insights into Broadway itself and beyond such as the impact of the Broadway Theatre on the economy of New York City [3]. Specifically, the problem we plan to investigate 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. We aim to provide descriptive analysis concerning the shows and predictive analysis that could help Broadway to better structure their shows in order to provide higher-quality service and to gain more profits.

2 Potential Analyzes that Can Be Conducted Using Collected Data

The data we plan to collect are as follows:

- ***Broadway Grosses*** for all the shows from 1985 to 2019, collected from http://www.playbill.com/grosses; (numerical)
- ***Broadway Social Stats*** for all the shows from 1985 to 2019, including both ratings and textual reviews, collected from https://www.broadwayworld.com/industry-social.cfm; (numerical)
- ***Broadway Reviews*** for all the shows from 1985 to 2019, collected from https://www.broadwayworld.com/reviews.cfm; (numerical & text)
- ***Broadway News*** for all the shows from 1985 to 2019, collected from http://www.playbill.com/news, the theatre-lover's go-to source for the latest news, interviews, photos, videos and more; (text)
- ***Broadway Shows' Wikipedia***, our supplementary data; (text)

To clarify, the data sets *Broadway Grosses* and *Broadway Social Stats* are considered as our primary numerical data and *Broadway Reviews* and *Broadway News* are considered as our primary text data, though part of the review data also have ratings, which are numerical. Section 3 provides an overview of each data set we will use for this project including how many records and attributes for each data set. As for the Broadway Shows' Wikipedia pages, we might not use them in the project, but since we will not have any opportunity to collect data later, we just collect them at this stage.

These data are meaningful for our data science problem in that they can tell different stories of a given show at Broadway. For instance, *Broadway Grosses* could demonstrate the financial aspect; *Broadway Social Stats* could help us build connections between the general public and people's reaction towards these shows. As a result, with all these data, we should be in a good position to investigate the the factors that contribute to a show's popularity and perform further analysis on social, economic, and cultural impact of Broadway shows.

Table 1 provides a list of selected variables in different data sets we have collected for illustration purpose. As mentioned above, these variables are

useful as they portrait different aspects of the shows. To be specific, for the *Broadway Grosses* data set, the variables *weekly_grosses* and *gross_difference* can tell us a show's performance in the market in a continuous and dynamic way. Similarly, the variable *attendance (seats_sold)* in this data set can tell us a show's popularity from audience's perspective. Likewise, variables in the *Broadway Social Stat* provide us information on certain shows from different social media platforms, which is important with regard to audience's reactions. In addition, textual data such as *reviews* could help us analyze what people think in a more detailed way. For instance, the use of strong positive or negative adjectives or the use of certain abbreviations could provide more information on people's sentiments towards certain shows.

Broadway Grosses (numerical)	***Broadway Social Stats (numerical)***	***Broadway Reviews & News (numerical & text)***	***Broadway Shows' Wikipedia (text)***
Weekly Grosses;	Facebook Likes;		Introduction;
Prev Week Gross;	Likes vs. Last Week;	Ratings;	Background;
Gross Difference;	Facebook Checkins;	Textual Reviews;	Plot Summary;
Average Ticket Price;	Twitter Followers;	Textual News;	Cast;
Attendance (seats_sold);	Instagram Followers;		Music;

Table 1: A List of Selected Variables in Different Types of Data Sets.

With all these data sets and their associated variables, there are several possible directions we may be able to investigate:

- Sale Predictions
  - This includes the sale performance of a given shown in the market so that Broadway can use such information to schedule different shows in order to meet audience's need as well as maximizing the profit.
- Seasonal/Influential Trend
- Annual/Seasonal Grosses

3 Data Issues

3.1 Numerical Data

Having explored our collected numerical data, we report the following data quality issues for each data set:

- ***Broadway Grosses***
  For dollar-amount attributes, there are dollar signs ($) preceding the numerical values. For attributes which represent percentage, there are percent signs (%) at the end of numerical values. To standardize the numerical data, it is necessary to remove those symbols.

  DATA TYPES: The ideal data type we consider for each attribute is demonstrated in Table 2:

Data Types	**int**	**string**	**float**
Attributes	seats_sold, perfs,	this_week_gross	diff_in_dollars, avg_ticket_price, percent_of_cap, different_percent_of_cap

Table 2: Data Types for Each Attribute in the Broadway Grosses Data Set.

  However, we have found that some of the numerical data are cast incorrectly into string values:

  DATA REDUNDANCY: For example, we have noticed that some data associated with 1985 are copies of those more recent data. It is possible that we have redundant data prevalent in our data sets, and thus it is of importance to check if the redundancy of the data and remove it accordingly.

  MISSING DATA & OUTLIERS: Those are data consisting of null or all zero values. It is worth noting that when a week is in the middle of a holiday such as Thanksgiving, theatres are closed so there are no grosses data associated with that week. We consider these data outliers instead of missing data (which need to be removed or replaced) since they reflect the seasonal nature of the Broadway business. For the shows which never have positive grosses in the entire data set, we consider them to be missing and need to be removed.

- ***Broadway Social Stats***
  Similar issues occur in *Broadway Social Stats* as in the *Broadway Grosses* data set: DATA TYPE, REDUNDANCY, and MISSING VALUE. Some columns contain extreme outliers, which mostly likely are erroneous data point recorded by the website.

  DATA TYPE: As mentioned above, this data set contains attributes such as "Facebook likes", "Facebook Checkins", "Instagram Followers", etc., and all of these should be numeric data. However, when we scraped

these data from the website, all of the value are in string format, thus further cleaning work should be applied to data types.

Data Redundancy: Intuitively, the record of each show in each week should only occur once. In other words, the combination of "the name of the show" and "the week of record" serves as the key to identify the only data record. However, some of the shows have more than one record in a week, which should be considered as redundant record.

Missing Data: Some of the values might be "space" or NA in the data set. We need to find them and process them properly. Besides, all the dates are missing year which can be misleading in further analysis. So, we need to infer the year of the date value by the sequence of the records.

Erroneous Data: Some of the outliers in the "Facebook likes" and "followers" data should be treated as errors. For example, if a Broadway show has X follower in week Y and then approximately the same amount of followers in week (Y+2), but less than 50% of followers than week Y in week (Y+1), these data points are very likely to be erroneous records as they are supposed to be relatively stable over weeks (without dramatic fluctuations).

- ***Broadway Reviews Ratings***
  The biggest issue with *Broadway Reviews Ratings* is that there are many missing values for shows. The *BroadwayWorld* website is founded in 2003, and as a result, if a show's first production on Broadway was earlier than this time, then it would not have a review page. Another possible reason is that if a show is a more recent or a new one, then it is likely that the reviews haven't been updated yet.

  Data Type: The data type has three different rating schema, "total ratings", "critics ratings", and "reader ratings", corresponding to the show names. The data type is corrected in this case.

  Data Redundancy: We checked the redundancy in this data set, and there is only one show that has been scrapped twice. This might be due to the fact that the list of show names have some duplicate copies.

  Missing Data: As mentioned above, some shows do not have any reviews due to two different reasons, that is, either the show is too old or too new. Some shows only have partial ratings and NA values. We

need to process this properly to preserve a balance between the volume of the data set and its integrity.

Erroneous Data: We noticed that some shows have total ratings and critic ratings as 0 in some cases. We checked the website and there are no reviews for this show yet. Thus, these are invalid ratings and we shall remove them afterwards.

3.2 Text Data

This section reports the issues we have found with our text data.

- ***Broadway Reviews***

  1. The textual review of a particular show is actually scrapped from the same page as numerical ratings. Thus, many shows do not have individual reviews from different news reports.
  2. Another problem is that the textual data contains "\n" to separate between paragraphs.
  3. When scrapping the title of a review, it actually contains more information including news company, author, and the date published, which requires further separation into more attributes.

- ***Broadway News***

  1. The news data have more comprehensive coverage than review data. We collected five latest news for each show from 1985-2019. There are only a few cases in which a show does not have five news entries.

- ***Broadway Shows' Wikipedia***

  1. Disambiguation seems to be a big issue in this data as some shows have generic names that could refer to several entries; as a result, shows with generic names cannot reach to the desired Wikipedia page as the rest of the shows.
  2. Some shows do not have a Wikipedia page, which results in missing data for those shows.

3. Depending on each show's structure, reputation, and popularity, the content is not formatted in the same way and does not necessarily have all the variables/attributes. For example, some shows may have a long history (e.g. *Cats*), then they would have more information included in their Wikipedia page as opposed to other shows with limited information.
4. Wikipedia pages are scraped as xml format, and thus they have many irrelevant tags such as *li, dv, td* etc. These tags should be removed as they are noises and not concerned in this analysis.

4 Collecting New Data

Table 3 provides an overview of the data set we intend to use in this project. For each data set, we list the number of records we have obtained and information about the attributes. Since *Broadway Social Stats* has 18 attributes, due to space limit, we do not list all of them in this table. We only select a few for illustration purpose.

	Broadway Grosses (numerical)	***Broadway Social Stats (numerical)***	***Broadway Review Ratings (numerical)***	***Broadway Reviews (text)***	***Broadway News (text)***
# Records	46,725	2,121	298	4,649	5,513
# Attributes	9	18	4	7	4
Attributes	week_ending, show, perfs, this_week_gross, diff_in_dollars, avg_ticket_price, seats_sold, percent_of_cap, diff_percent_of_cap	show, date, FB likes, Likes Vs.Last Week, FB Checkins, Twitter Followers etc.	show; total_rating; critics_rating; readers_rating	show, rating, title, content, company, author, date	show, title, subtitle, content

Table 3: An Overview of the Data Sets used in this Project.

4.1 Primary Data Set 1: Numerical Data

- ***Broadway Grosses***
  This data set consists of 46,929 records and 9 attributes:

  WEEK_ENDING: The time stamp of the week in question
  SHOW: The name of the Broadway show
  SEATS_SOLD: Number of seats sold this week
  PERFS: Number of performance in a week
  THIS_WEEK_GROSS: The gross of this week
  DIFF_IN_DOLLARS : Gross difference comparing to last week
  AVG_TICKET_PRICE: Average ticket price

PERCENT_OF_CAP: Percentage of seats sold out of total seats
DIFF_PERCENT_OF_CAP: Difference of percentage of seats sold out of total seats comparing to last week's

- ***Broadway Social Media Stats***
  This data set consists of 2,121 records and 18 attributes:

  As shown in Table 4, most of the attributes are self-explained in their names. To clarify two attributes:

  – The variable CURRENT has two variances: *current* or *upcoming*.
  – The variable TYPE indicates the type of the show: whether it is a musical or it is a play.

Show	Date	FB Likes	Likes Vs.Last Week	FB Talking About	Talking Vs.Last Week
FB Checkins	Checkins vs.Last Week	Twitter Followers	Twitter vs.Last Week	Instagram Followers	IG Followers vs.Last Week
Current	Type	Total Fans Change	Month	Day	Year

Table 4: A List of Attributes in the *Broadway Social Media Stats* Data Set.

- ***Broadway Review Ratings***
  This data set consists of 298 records and 4 attributes:

  SHOW: The name of the Broadway show
  TOTAL_RATING: The overall rating of the show
  CRITICS_RATING: The rating from the critics
  READERS_RATING: The rating from the readers

4.2 Primary Data Set 2: Text Data

- ***Broadway Reviews***
  This data set consists of 4,649 records and 7 attributes:

  SHOW: The name of the Broadway show
  RATING: The rating from a specific critic to the show
  TITLE: The title of the review
  CONTENT: The content of the review
  COMPANY: The authorization that completes the review
  AUTHOR: The author who wrote this review
  DATE: The date on which the review is published

- ***Broadway News***
  This data set consists of 5,513 records and 4 attributes:

  SHOW: The name of the Broadway show that has been searched
  TITLE: The title of the news
  SUBTITLES: The subtitle of the news, a brief summary of the content
  CONTENT: The content of the news

4.3 Supplementary Data: Wikipedia for Broadway Shows

There are 1,099 files (out of 1,110 shows) for the Broadway shows examined here (i.e. 1985-2019). In addition to removing some irrelevant tags in the source files, we have not uniformed the structure yet since we do not know what information we would like to include later in the project we just keep all the information we have obtained.

5 Data Cleanliness

- ***Broadway Grosses***

  We report the overall score for this data set as well as quality score for each attribute, which are shown in Table 5.

  Three quality metrics are calculated to measure the data quality of each attribute:

  Firstly, in terms of missing values, both null values and zero values are taken into account. More specifically, since zero grosses may be due to the fact that theaters are closed during the holiday seasons, we only consider records as missing values if the associated shows never have positive grosses in the entire data set. Based on those criteria, the missing value quality score reflects "1 - (# of missing values / # of all values)" and is converted into a scale of 1 to 100.

  Secondly, a data type metric is designed to measure if a certain attribute has the 'ideal' data type for future usage. A score of 100 indicates a correct data type while a score of 0 indicates an incorrect one.

  Thirdly, redundancy is determined based on show names, the grosses of this week, and the gross difference comparing to last week. Note

that records of zero values are also duplicates of themselves. Since the quality measurement of missing value has taken them into account, we do not count redundancy metric on those data again. An overall score is calculated for each attribute by averaging the scores of the three quality measures.

From Table 5 we can observe that attributes such as *this_week_grosses* and *percent_of_cap* result in relative low scores due to the incorrect data types. It is also worth noting that data issues associated with the *Broadway Grosses* data set are diluted by its large sample size. A small fraction of missing and redundant data in a data set of over 40,000 records would not make a big difference, as opposed to the same amount of poor quality data in a smaller data set.

Quality Scores before Data Cleaning				
	missing_value	***wrong_data_type***	***redundancy***	***Overall Score***
week_ending	100	100	99.853	99.951
show	99.998	100	99.853	99.95
this_week_gross	99.847	0	99.853	66.567
diff_in_dollars	99.847	0	99.853	66.567
avg_ticket_price	99.847	0	99.853	66.567
seats_sold	99.847	100	99.853	99.9
perfs	99.847	100	99.853	99.9
percent_of_cap	99.847	0	99.853	66.567
diff_percent_of_cap	99.847	0	99.853	66.567
Quality Scores after Data Cleaning				
	missing_value	***wrong_data_type***	***redundancy***	***Overall Score***
week_ending	100	100	100	100
show	100	100	100	100
this_week_gross	100	100	100	100
diff_in_dollars	100	100	100	100
avg_ticket_price	100	100	100	100
seats_sold	100	100	100	100
perfs	100	100	100	100
percent_of_cap	100	100	100	100
diff_percent_of_cap	100	100	100	100

Table 5: Overall Quality Score and Quality Score for Each Attribute in the *Broadway Grosses* Data Set before and after Data Cleaning.

- ***Broadway Social Media Stats***

Before Data Cleaning			***After Data Cleaning***	
Show	99.93		Show	100
Date	99.93		Date	100
FB Likes	74.43		FB Likes	75
Likes Vs. Last Week	74.93		Likes Vs. Last Week	100
FB Talking About	74.93		FB Talking About	100
Talking Vs. Last Week	74.93		Talking Vs. Last Week	100
FB Checkins	69.43		FB Checkins	74.75
Checkins vs. Last Week	74.93		Checkins vs. Last Week	100
Twitter Followers	70.68		Twitter Followers	75
Twitter vs. Last Week	74.93		Twitter vs. Last Week	100
Instagram Followers	73.43		Instagram Followers	75
IG Followers vs. Last Week	74.93		IG Followers vs.Last Week	100
Current	99.93		Current	100
Type	99.93		Type	100
Total Fans Change	74.93		Total Fans Change	100
Overall Score	**80.81**		***Overall Score***	**93.32**

Table 6: Overall Quality Score and Quality Score for Each Attribute in the *Broadway Social Media Stats* Data Set before and after Data Cleaning.

For this part, we take 4 dimensions into consideration as mentioned above: Data Types, Data Redundancy, Missing Values, and Erroneous Data. The data quality metric is designed to evaluate this data set. Specifically, the mean score of the four dimension is the total score of the data set. For each dimension, the score is defined as "1 - percentage of error" of the data points. The scores are converted to a scale of 1 to 100, the quality score of each attribute is assigned as the average score of all the dimensions. The overall quality score of data set is 80.81 before and 93.32 after data cleaning. We report the overall score for this data set as well as quality score for each attribute, which are shown in Table 6. As can be seen in the table, most of the attributes did not receive a satisfying score. Most of the numeric attributes were in the string format when scraped from the website. Also, there were redundant data that we obtained more than one record during the same period of the same show. For Facebook, Twitter and Instagram related data, erroneous data points are observed as we discussed above.

- ***Broadway Reviews Ratings***

Missing data are prevalent in this data set since not many shows have individual reviews as well as ratings. Therefore, we primarily use the missing data and redundancy metrics to measure its data quality. We report the overall score for this data set as well as quality score for each attribute, which are shown in Table 7.

Quality Scores before Data Cleaning			
	missing_value	***redundancy***	***Overall Score***
show	100	99.909	99.954
total_rating	27.157	99.909	63.533
critics_rating	27.157	99.909	63.533
readers_rating	26.703	99.909	63.306
Quality Scores after Data Cleaning			
	missing_value	***redundancy***	***Overall Score***
show	100	100	100
total_rating	27.091	100	63.546
critics_rating	27.091	100	63.546
readers_rating	26.626	100	63.318

Table 7: Overall Quality Score and Quality Score for Each Attribute in the *Broadway Review Ratings* Data Set before and after Data Cleaning.

6 Data Cleaning

- ***Broadway Grosses***

  The data cleaning procedure mainly consists of:

  1) standardizing numerical data,
  2) removing missing data, and
  3) removing redundant data

  In terms of numerical data standardization, we eliminates special symbols such as $, which is associated with any dollar amounts, and %, which is associated with any percentage values. These symbols reduce the usability of quantitative analysis, and their presence results in an incorrect data type categorization of the numerical values. After re-

moving them and converting numerical values from strings to float data types, performing statistical summary on the data becomes possible.

Missing data are determined based on two criteria. One is when a record consists of null values, and the other one is when a row is full of zero values. After exploring the data, we have observed that, while some of the zero values are the indication of missing data, other are due to holiday seasons when theaters are closed and no Broadway shows are being performed. The latter would be valuable for our analysis since it reflects the seasonal nature of the Broadway industry. Therefore, we only remove null values and records of shows with only zero grosses and statistics.

For the redundancy cleaning, we compare records by their show names, grosses of this week, and the gross difference comparing to last week. Records of the first occurrence of the shows are kept and the following duplicates are removed. Note that the records of 1985-06-02 are manually deleted. We have noticed that it is a copy of data of 2019-09-26, but since we did not include the data of 2019-09-26 in this data set, there is no way to use an automated process to make comparisons and remove the redundant data.

The cleaned data set resulted from the cleaning procedure above have 9 attributes and 46,725 rows of data records. The quality measure program also gives it 100 scores in terms of the three data quality dimensions.

- ***Broadway Social Media Stats***

  The data set is well-cleaned after the data cleaning process. We removed the redundant values, changed the incorrect data types, and modified the outliers. The most important two data cleaning steps (towards attributes) are described below:

  Erroneous Data: There are some erroneous data points in the data set, leading to the most serious issue in the *Broadway Social Media Stats* data set. The issue can be observed when the record of a show drops or increases dramatically in one week, but seems normal a week before and a week after. For example, a show has 50,000 followers in Twitter in both Week 1 and week 3, but the number of followers dropped to only 5,000 in Week 2, which is very likely to be an error.

This type of error occurred in 4 attributes: FACEBOOK LIKES and FACEBOOK CHECKINS, which intuitively should be gradually increase over time; TWITTER FOLLOWERS and INSTAGRAM FOLLOWERS, which should not have dramatic fluctuations.

To deal with this issue, we identified the outliers with a standard of "less than 50% of the neighboring weeks" or "larger than 150% of the neighboring weeks", and assigned the mean of the neighboring week values to them.

After data cleaning, the score of erroneous data reached 99.99 for FACEBOOK LIKES, FACEBOOK CHECKINS, TWITTER FOLLOWERS, and INSTAGRAM FOLLOWERS. The final scores of these 4 attributes are around 75 because they still have low score for the data type score (i.e. the type was changed from *string* to *int*, but read as *float* when being evaluated). The total score of the *Broadway Social Media Stats* data set has been improved from 81 to 93.

MISSING VALUE: Although the data set itself does not contain missing values, the information on the year of records is missing. Hence, it is necessary to infer the year of the date. The data set is sequenced by time, so the upper ones are more recent. To clean the data, we compare the month value of each records within each show. If the month of a record is smaller than the previous row (and they have the same "show name"), this can help us determine that it is from the year before.

- ***Broadway Review Ratings***

  We excluded the records whose three ratings are NAs or 0 since these are due to the errors from the website's part. They do not contribute meaningful analysis on the reputation of the shows. We also dropped one row that is a duplicate of the other one. We kept the records with at least one valid rating, which could further be correlated with other textual review data.

- ***Broadway Textual Reviews***

  As mentioned in the Section 3.2, the whitespaces are represented by "\n". We striped the paragraph before appending it to the data frame. We also extracted information COMPANY, AUTHOR, and DATE from the TITLE attribute and added them as new columns to the data set.

References

[1] Yaakov Bressler. Opening the Stage Door for Big Data in Broadway – Building Databases from Unstructured Text using Machine Learning, Sep 2018.

[2] Michael Gold, Ryan McClarren, and Conor Gaughan. The Lessons Oscar Taught Us: Data Science and Media & Entertainment. *Big data*, 1(2):105–109, 2013.

[3] Mathtech. The Impact of the Broadway Theatre on the Economy of New York City, May 2019.