Datasheet for Broadway Gross Dataset*

The dataset used for 'Understanding Broadway Gross: Key Factors and Growth Strategie'

Xuanle Zhou

December 2, 2024

This datasheet is the extract of the questions from Gebru et al. (2021). And it was put together with the help of Alexander (2023)

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - The dataset was created to enable the analysis of weekly gross revenue for Broadway productions. While it does not address a specific gap, it provides a structured repository of gross data that is valuable for understanding industry trends and evaluating show performance.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The specific creators of the dataset have not been explicitly identified. However, it is likely associated with The Broadway League and their Internet Broadway Database (IBDB) (Database 2024), the official archive for Broadway theatre information. IBDB serves as a comprehensive resource for historical and current Broadway data.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - Information regarding the funding source for the creation of the dataset is not currently available. However, The Broadway League operates in partnership with entities such as Chase and The New York Times, which may have contributed to its initiatives.

4. *Any other comments?*

*Code and data are available at: https://github.com/Isazhou13/Broadway_Gross

- No further comments at this time.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
    - Each instance represents the information for a single production during a specific week.
2. *How many instances are there in total (of each type, if appropriate)?*
    - The dataset contains a total of 47,524 instances.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
    - The dataset includes all available instances of weekly gross data for Broadway productions in New York City, as collected by The Broadway League.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

The raw data and the accompanying data dictionary are provided by Cookson (2020).

- week_ending: Date of the end of the weekly measurement period. Always a Sunday. -week_number: Week number in the Broadway season. The season starts after the Tony Awards, held in early June. Some seasons have 53 weeks.
- weekly_gross_overall: Weekly box office gross for all shows
- show: Name of show. Some shows have the same name, but multiple runs.
- theatre: Name of theatre. Only shows most recent theatre for shows that started at one theatre and moved to another (e.g., The Lion King will show Minskoff Theatre even though it played at New Amsterdam Theatre from 1997-2006).
- weekly_gross: Weekly box office gross for individual show
- potential_gross: Weekly box office gross if all seats are sold at full price. Shows can exceed their potential gross by selling premium tickets and/or standing room tickets.
- avg_ticket_pric: Average price of tickets sold
- top_ticket_price: Highest price of tickets sold
- seats_sold: Total seats sold for all performances and previews
- seats_in_theatre: Theatre seat capacity

- pct_capacity: Percent of theatre capacity sold. Shows can exceed 100% capacity by selling standing room tickets.
- performances: Number of performances in the week
- previews: Number of preview performances in the week. Previews occur before a show's official open.

5. *Is there a label or target associated with each instance? If so, please provide a description.*
    - No specific label or target is explicitly assigned to each instance
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
    - Some information is missing from individual instances. Specifically, data on potential gross is unavailable for dates prior to July 1997. This is because Broadway productions did not calculate or record this information before that time.
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*
    - No explicit relationship available
8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
    - No recommended data splits available
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
    - There is no errors, sources of noise, or redundancies in the dataset
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - The dataset is primarily self-contained but may reference external resources like theater locations or production details available via the Internet Broadway Database (IBDB). These resources are managed by The Broadway League, and their availability over time cannot be guaranteed without archival support.

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - No

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - No

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

    - No

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - No

16. *Any other comments?*

    - No further comments

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - The data is directly collected through observable metrics.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

- The specifics of this process have not been disclosed.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - The dataset represents the complete set of data and is not a sample from a larger collection.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - Staff from organizations, including theater managers, producers, and accountants, are responsible for the data collection process. They are paid by salary.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

    - According to The Broadway League, they began recording weekly gross data in 1979. However, it is unclear whether they recorded the additional information found in the dataset during that time. The dataset contains detailed weekly gross data starting from the first week of July 1985.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

    - I obtained the data via third parties. I downloaded the data from a publicly available GitHub repository provided by Cookson (2020), which sourced the information from the Playbill website (Playbill 2024). The link for the GitHub is broadway_grosses. The R Core Team (2023) package is used for downloading the data.

8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

    - For tickets purchased online, users are presented with terms and conditions when creating an account. However, it remains unclear how information from previous years was collected or whether individuals were notified during that process.

9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*

    - Consent is inherently granted through the membership agreement with The Broadway League. By accepting the terms of membership, individuals agree to submit their data and allow its use for industry analysis and reporting.

10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

    - Yes, the privacy policy states: "We may share your Personal Information with third parties for third-party marketing purposes. You can opt out of this by emailing us at league@broadway.org."

11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

    - No.

12. *Any other comments?*

    - No further comments

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

    - Yes, the data cleaning for this study is complete. The raw data included a week_ending column that combined the year, month, and date into a single cell. To address this, the code was used to separate these elements into distinct columns. Unnecessary columns for this study, such as `weekly_gross_overall`, `show`, `theatre`, `potential_gross`, `top_ticket_price`, `seats_sold`, `pct_capacity`, and `previews`, were removed. A new column, `holiday_week`, was created, with values marked as 1 for weeks corresponding to Independence Day, Labor Day, Thanksgiving, Christmas, and New Year's Week. Additionally, a `Tony_Awards` column was constructed to indicate whether a given month corresponds to the Tony Awards period. The Tony Awards is a prestigious theater award typically held in June. Moreover, the data is limited to the most recent decade, specifically from 2010 to 2020. Unreasonable values were filtered out, such as shows with an average ticket price of 0.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
    - Yes, the raw data is saved. It can also be accessed at this link.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - The software used is R Programming Language (R Core Team 2023).
4. *Any other comments?*
    - No further comments

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - Yes, the dataset is publicly available and has been used in many papers for analysis.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
    - No.
3. *What (other) tasks could the dataset be used for?*
    - Some tasks using the dataset involve more detailed analyses of gross changes for particular shows.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
    - Since the data is filtered from 2010 to 2020, it is recommended to evaluate whether this time frame is appropriate for future detailed analyses.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
    - No, there are no tasks for which the dataset should not be used.
6. *Any other comments?*
    - No further comments.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
    - The dataset is publicly available through GitHub repositories.
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - The dataset is available from GitHub, following link of: https://github.com/tacookson/data/tree/mas grosses
3. *When will the dataset be distributed?*
    - The dataset was published in April 2020.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - No.
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - No.
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - No.
7. *Any other comments?*
    - No further comments.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The GitHub repository owner or contributors hosting the dataset typically maintain it, with the original data provided by The Broadway League.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

- You can contact the owner by creating an issue on the GitHub repository. Additionally, the owner's website homepage includes contact information. The owner's email is alexander.cookson@gmail.com, and their LinkedIn profile can be accessed at https://www.linkedin.com/in/alexcookson/.

3. *Is there an erratum? If so, please provide a link or other access point.*

   - No.

4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*

   - It does not provide information regarding this issue.

5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*

   - Retention limits are determined by The Broadway League's policies.

6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*

   - There is no information regarding this issue.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*

   - They can contact the Internet Broadway Database using the form available here. This form can be used to contact Broadway with questions or to provide additional information. However, any updates made may only be reflected in the Broadway database and not in the GitHub repository.

8. *Any other comments?*

   - No further comments.

References

Alexander, Rohan. 2023. *Telling Stories with Data*. "University of Toronto". https://www.tellingstorieswithdata.com.

Cookson, Tim. 2020. "Broadway Grosses Dataset." https://github.com/tacookson/data/tree/master/broadway-grosses.

Database, Internet Broadway. 2024. "Statistics." https://www.ibdb.com/statistics/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Playbill. 2024. "Grosses List." https://playbill.com/grosses.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.