Movies Dataset

Motivation:

The reason this dataset was created was for data analysis practice. Given its variables, it could be used to estimate what variables make a well-rated movie or how well upcoming movies might perform at the box office.

Dataset Composition:

There are 3907 instances in the original movie dataset. Each instance contains the following attributes: movie, director, running time, top three actors, genre, budget, box office total, actor's box office percentage, director’s box office percentage, earnings, Oscar and Golden Globes nominations, Oscar and Golden Globes awards, release year, and IMDb score. Everything is included in the original dataset and no initial experiments were carried out.

Data Collection Process:

A human (user Delfina Oliva on Kaggle) collected the data from IMDb, Wikipedia, and The Movie Database (TMBD), which was directly observable. It does not contain all possible instances, but rather a random selection of movies. Some attributes contain the Unicode symbol “�” for certain letters that were not imported correctly.

Data Processing:

The instances containing the “�” symbol were decoded back to their original letters, underscores were replaced with spaces in the column names, and the budget and box office total were changed from scientific to decimal format. The original data was saved in the same directory alongside the edited version, and the processing software used to clean the data was RStudio.

Dataset Distribution and Maintenance:

The data will be distributed on GitHub with no fees or restrictions. The host of this dataset is the Kaggle user who created it, and they plan to update it annually.

Word Bank Dataset

Motivation:

This dataset was created to understand children’s comprehension and language development across different languages, ages, health conditions, etc. This dataset has been used for eight research papers already, however, there may be more that are not directly listed on the Wordbank website. The development of Wordbank was funded by a grant from the National Science Foundation (#1451577) as well as the MB-CDI Advisory Board.

Dataset Composition:

There are 14,826 instances in the selected section of the dataset I chose, and each instance represents a child. The attributes in the selected section are language, age, comprehension, production, is norming, birth order, caregiver education, race, sex, health conditions, monolingual, and typically developing. Everything is included in the dataset without relying on external sources and there were no initial experiments run on this dataset.

Data Collection Process:

The data is collected by an ever-growing group of researchers who have sent their findings into Wordbank so that they can convert it into their “Wordbank format.” However, since every researcher collects data on their own it is difficult to pinpoint exactly where they source their data from.

Data Processing:

The raw data was saved alongside the cleaned version of the dataset using RStudio. The original creators of the data use a MySQL database managed using Python and Django. This processing allows for its initial motivation behind creation because by using different tools given by Wordbank anyone can create a dataset using their vast databases to suit their research needs.

Dataset Distribution:

WordBank is hosted on many different platforms. There is a website (<https://wordbank.stanford.edu/data/?name=admin_data>) hosting databases that can be customized, and they have an API that can be accessed through R through the wordbankr package. Also, all the data can be downloaded through the website itself or their GitHub repository without any fees or restrictions. Wordbank and its datasets are licensed under a Creative Commons Attribution 4.0 International License, [CC BY 4.0](https://creativecommons.org/licenses/by/4.0/) and Some datasets listed under the [Contributors page](http://wordbank.stanford.edu/contributors) may be marked with a CC BY NC License 4.0logo meaning these datasets are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, [CC BY-NC 4.0](https://creativecommons.org/licenses/by-nc/4.0/).

Dataset Maintenance:

Wordbank maintains the dataset which is updated as often as researchers contribute to the database.

Legal & Ethical Considerations:

All the data in Wordbank is completely anonymized an does not constitute human subjects research. “Potentially identifying information such as date of birth is used to compute child’s age of test, but is never stored in the database. We assume that your data were collected following your own institution’s IRB standards and should be completely anonymized before contributing to Wordbank.”