Project ETL

Team 4: Joe Adams, Michelle Lucio, Ben Farniok, and Jessica Schmitz.

**SUMMARY:**

**INITIAL DATA SETS**

Here are links to the original IMDB data sets:

<https://datasets.imdbws.com/title.akas.tsv.gz>

<https://datasets.imdbws.com/name.basics.tsv.gz>

<https://datasets.imdbws.com/title.basics.tsv.gz>

<https://datasets.imdbws.com/title.ratings.tsv.gz>

<https://datasets.imdbws.com/title.principals.tsv.gz>

**CLEANING THE DATA**

* <https://datasets.imdbws.com/title.akas.tsv.gz>
  + Decided that this data wasn’t useful and ended up not using IT
* <https://datasets.imdbws.com/title.basics.tsv.gz>
  + Removed inappropriate content (adult films –“isAdult” column indicates this content)
    - Once adult films are removed from the database, we then removed the column “isAdult” as it wasn’t necessary since the remaining values are all not adult content
  + Renamed columns
    - ttconst to titleID
    - titelType to Type
  + set index based on “titleID”
  + Replace \N to N/A
* <https://datasets.imdbws.com/title.principals.tsv.gz>
  + Removed columns
    - Ordering
    - job
  + Renamed columns
    - ttconst to titleID
    - nconst to nameID
  + Replace \N to N/A
* <https://datasets.imdbws.com/title.ratings.tsv.gz>
  + Renamed columns
    - ttconst to titleID
* <https://datasets.imdbws.com/name.basics.tsv.gz>
  + Removed columns:
    - birth year
    - death year
  + Replace
    - Replaced \N with N/A then dropped N/A values in “assocTitlesID” column
  + Renamed columns
    - nconst to nameID
    - knowforTItles to assocTitlesID

**CREATING THE DATA BASE**