Document for Group Project: The Office

Project Goal: To see if there is a correlation between the amount a character speaks with the IMDB rating.

Extracted Data Source:

* The extraction was from Kaggle
  + The Office IMDB Per Episode
  + The Office Lines
* The Office IMDB Per Episode data set has the following columns: season, title, air date, rating, number of votes. Description of the episode, directed by, and written by.
  + Extracted the following columns: season, title, and rating.
  + This will give us a joining to include the rating of each episode.
* The Office Lines data set has the following columns: season, episode, title, speaker, and line.
  + Extracted the following columns: season, title, episode, and speaker.
  + This will provide the joining with either season or title with the speaker.

Transformation:

* At face value the data looked cleaned and easy to use. However, once we studied the data we discovered it was not the case.
  + Misspelling are throughout The Office Lines data set for the speakers and the title.
    - We cleaned by replacing the misspelled name with the correct name.
    - We cleaned the title in the same manner with the speaker name.
  + Joined the data sets with title columns with a left join.

Relational or Non-Relational Database:

The final tables or collections are used in the production database: