***Data Collection:***

**Source**: Reddit

(<https://docs.google.com/spreadsheets/d/18wS5AAwOh8QO95RwHLS95POmSNKA2jjzdt0phrxeAE0/edit#gid=747974534>)

**File Type**: Google Sheet

**Data:** Provided each actors line per Season and episode. Also provided a column if the line came from a deleted scene or not.

**Source**: IMDB

<https://www.imdb.com/list/ls000977475/?sort=release_date,asc&st_dt=&mode=detail&page=1>

<https://www.imdb.com/list/ls000977475/?st_dt=&mode=detail&page=2&sort=release_date,asc>

**File Type**: HTML

**Data**: Provided a rating and a synopsis for each episode from The Office.

***Data Scraping/Cleaning:***

**Programming Language**: Python

* Via Jupyter Notebook we used dependencies like Beautiful\_Soup in order to scrape the HTML files and pandas to read the google sheet in order to transform them into workable DataFrames.

Our QA process in the Google Sheet and the IMDB pages:

* Confirming data types and its data were uniform across all column fields
* Confirming there weren’t no misspelling in actor’s name.
* Removing columns that weren’t necessary for our project

***Database Setup:***

Relational Database: SQLite

We imported the DataFrames from Pandas directly to SQLite

We created two different tables:

**Lines:** This table provides the line count per actor in each season and episodes.

**Metadata:** This table provides season/episode ID, title of the episode, synopsis of the episode, IMDB rating and episode number.

***Application Development:***

Flask Application

JavaScript File

Heroku Deployment