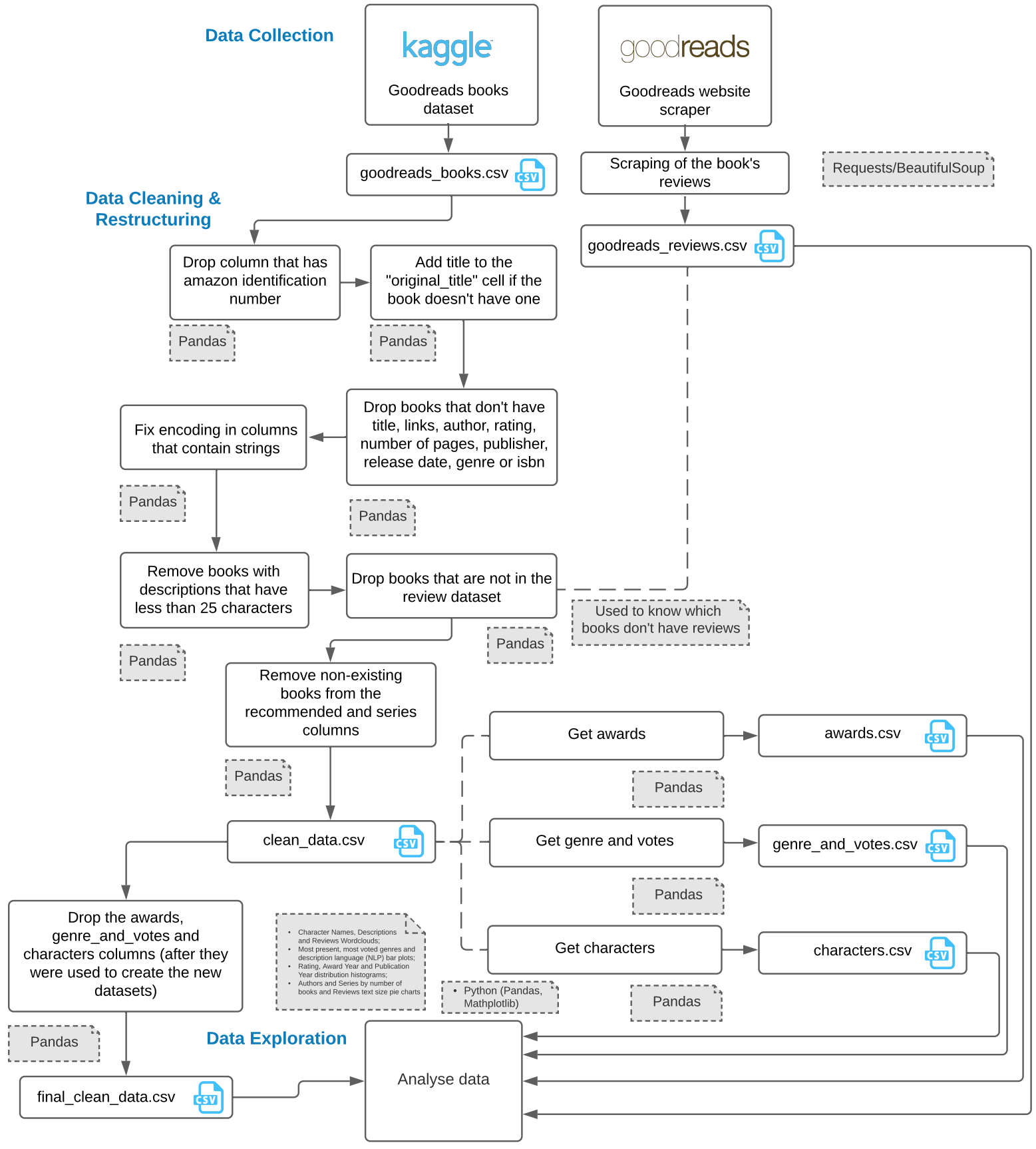
Data Collection
kaggle
Goodreads books dataset
goodreads
Goodreads website scraper
goodreads_books.csv
Scraping of the book's reviews
Requests/BeautifulSoup
Data Cleaning & Restructuring
goodreads_reviews.csv
Drop column that has amazon identification number
Add title to the "original_title" cell if the book doesn't have one
Pandas
Pandas
Drop books that don't have title, links, author, rating, number of pages, publisher, release date, genre or isbn
Fix encoding in columns that contain strings
Pandas
Pandas
Remove books with descriptions that have less than 25 characters
Drop books that are not in the review dataset
Used to know which books don't have reviews
Pandas
Pandas
Remove non-existing books from the recommended and series columns
Pandas
Get awards
awards.csv
Pandas
clean_data.csv
Get genre and votes
genre_and_votes.csv
Pandas
Drop the awards, genre_and_votes and characters columns (after they were used to create the new datasets)
Character Names, Descriptions and Reviews Wordclouds;
Most present, most voted genres and description language (NLP) bar plots;
Rating, Award Year and Publication Year distribution histograms;
Authors and Series by number of books and Reviews text size pie charts
Get characters
characters.csv
Python (Pandas, Mathplotlib)
Pandas
Pandas
Data Exploration
final_clean_data.csv
Analyse data