

Pitchfork Music Reviews, Milestone #1, Information Processing and Retrieval

Fabio Huang - up201806829,

Lisa Sonck - up202202272,

Temaco - up202200606

Content



- Data Collection
- Original Dataset
- Data Cleaning
- Data Characterization
- Conceptual Model
- Data Processing Pipeline
- Queries and Search Tasks

Data Collection

The Kaggle logo, featuring the word "kaggle" in a light blue, lowercase, sans-serif font.

18,393 Pitchfork Reviews

Pitchfork reviews from Jan 5, 1999 to Jan 8, 2017



Original Dataset

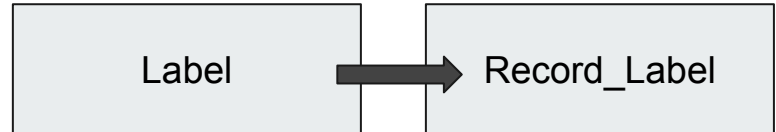
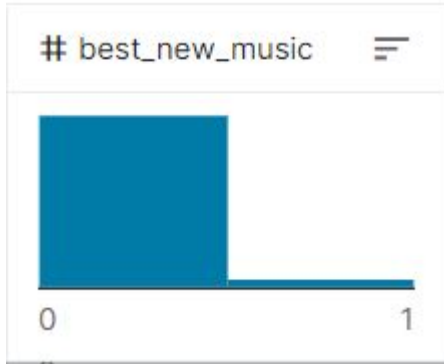
- Merging all the tables together

Table	Total Rows
artists	18831
content	18393
genres	22680
labels	20190
reviews	18393
years	19108

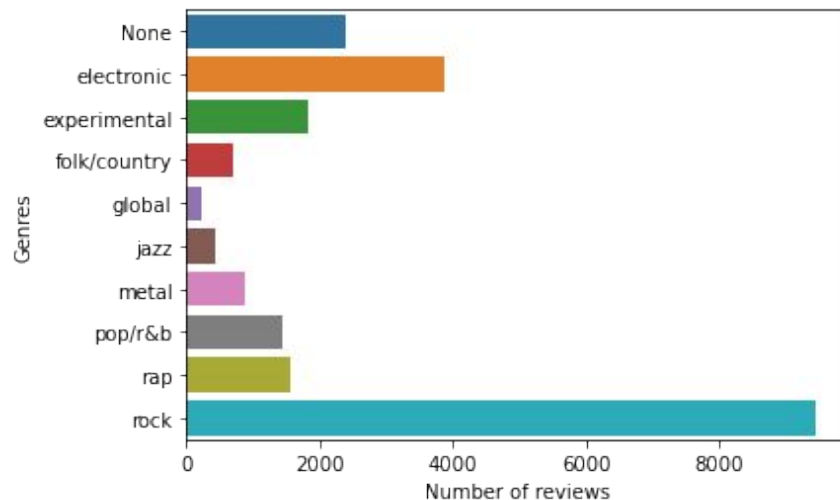
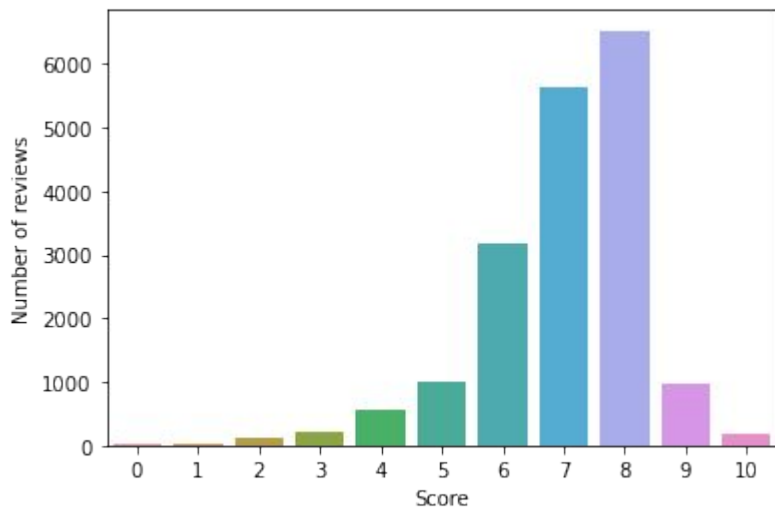
	reviewid	title	artist_x	url	score	best_new_music	author	author_type	pub_date	pub_weekday	pub_month	pub_year	genre	year	label	artist_y	content
0	22703	mezzanine	massive attack	http://pitchfork.com/reviews/albums/22703-mezz...	9.3	0	nate patrin	contributor	2017-01-08	6	1	2017	electronic	1998	virgin	massive attack	"Trip-hop" eventually became a '90s punchline,...
1	22721	prelapsarian	krallice	http://pitchfork.com/reviews/albums/22721-prel...	7.9	0	zoe camp	contributor	2017-01-07	5	1	2017	metal	2016	hathenter	krallice	Eight years, five albums, and two EPs in, the ...
2	22659	all of them naturals	uranium club	http://pitchfork.com/reviews/albums/22659-all-...	7.3	0	david glickman	contributor	2017-01-07	5	1	2017	rock	2016	shock,fashionable idiots	uranium club	Minneapolis' Uranium Club seem to revel in bei...
3	22661	first songs	kleenex, liliput	http://pitchfork.com/reviews/albums/22661-firs...	9.0	1	jenn pelly	associate reviews editor	2017-01-06	4	1	2017	rock	2016	kill rock stars,mississippi	kleenex,liliput	Kleenex began with a crash. It transpired one ...
4	22725	new start	taso	http://pitchfork.com/reviews/albums/22725-new-...	8.1	0	kevin lozano	tracks coordinator	2017-01-06	4	1	2017	electronic	2016	teklife	taso	It is impossible to consider a given release b...

Data Cleaning

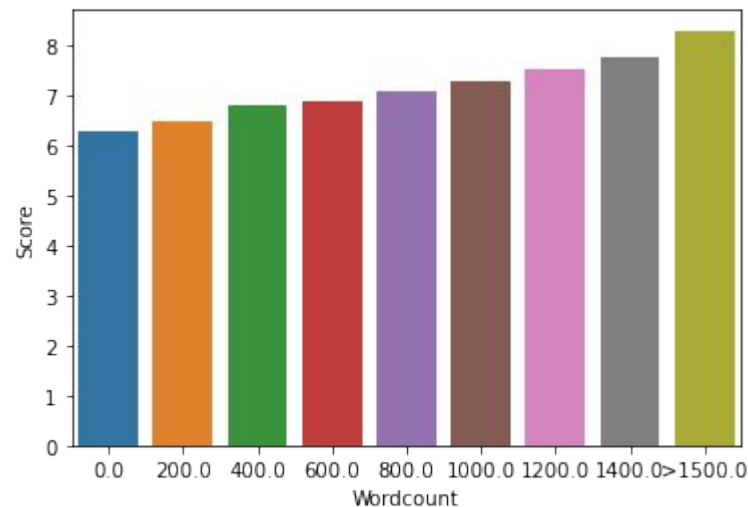
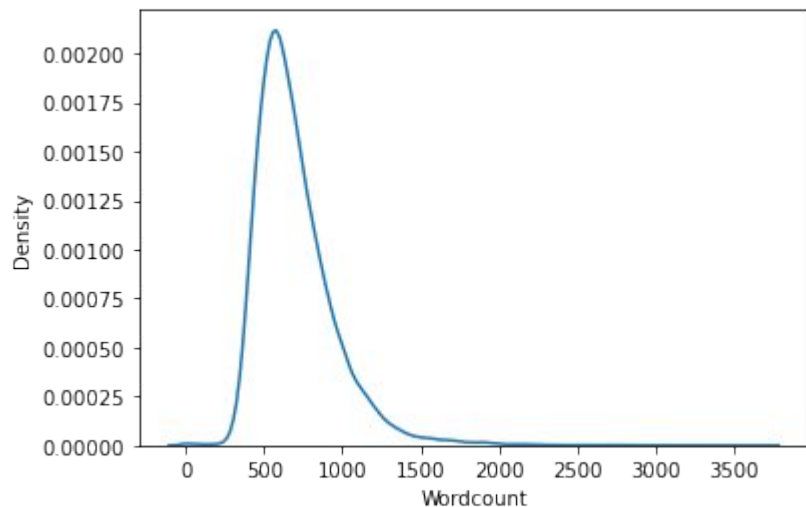
- Drop irrelevant and duplicated columns
- Removed entries with empty 'review_content'
- Renamed a few columns



Data Characterization



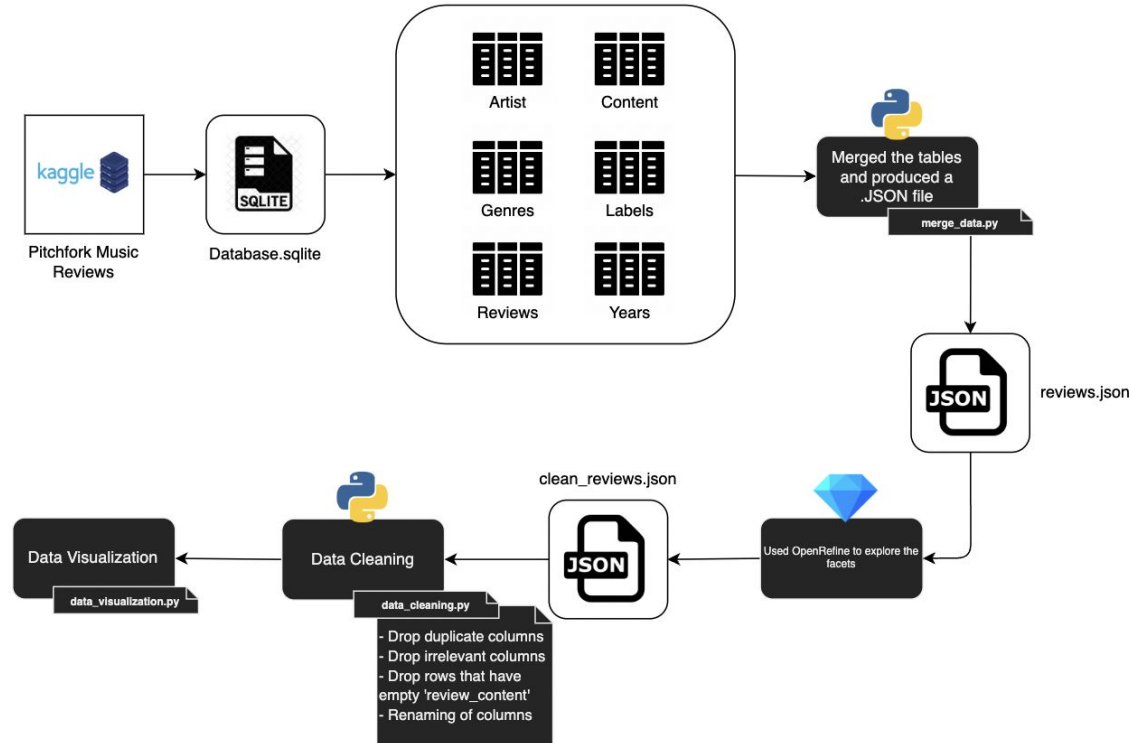
Data Characterization



Conceptual Model



Data Processing Pipeline



Queries and Search Tasks



- **Search for a review**
 - based on the **authors name**, the **genre**, the **score** or the **content of the review**.
- **Search for a song**
 - based **genre**, **artist**, **review-score**, **name**.
- **Check which genres are more reviewed by Pitchfork.**
- **Check which authors write good reviews per genre.**
 - Returns a list of authors who have written more than 10 reviews for one genre of music, where all these reviews have a total average score of eight.

References



- [1] Nolan Conaway. 18,393 Pitchfork Reviews.
<https://www.kaggle.com/datasets/nolanbconaway/pitchfork-data>