**Introduction**

Since 1952, *Sight and Sound* (SaS)—a publication nested within the British Film Institute (BFI)—has conducted a decennial poll asking critics, programmers, curators, archivists, academics, and other industry professionals to each select ten great films. These selections are then tallied, and a list of the ‘greatest’ films is compiled. Per BFI, this list is the “major bellwether” of critical opinion on cinema; the 2022 (and eighth) edition of the list, with 1,639 voters, is also the largest ever.

Lists are weird cultural objects, but the BFI’s “Greatest Films of All Time” does form a loose and informal ‘canon’ that’s treated with, I believe, both reverence and suspicion. Transforming the list into a dataset offers researchers—primarily in digital humanities, film studies, and cultural studies—a valuable reference point for what film workers, in the years 2012-2022, thought to be valuable.

**Preproduction** 📝

My dataset’s scaffolding has already been established by a GitHub user. I’ve extracted data from this user’s work that will be most valuable to my project. I’m excluding some variables from my new dataset that were originally collected (e.g. Director Ratings, Voter Profiles) that are fantastic and could be valuable to different types of analysis. To find a link to this user’s work, please see this project’s README.

Note: In the preexisting dataset, there are far more observations than in my modified list, as even films with a single vote in SaS’s database have been captured by the Python program the user wrote. My dataset only focuses on the list of films that SaS has published and presented to users on their website.

**Production** 🎬

For enrichments, I’ve relied on IMDb, which compiles vast amounts of information about films from various sources, including credits, pressers, biographies, interviews, and user contributions. Though IMDb data goes through consistency checks, it’s impossible to validate every piece of metadata attached to a film or director. In moments of doubt, I’ve supplemented my research using Wikipedia.

‘Genre’ as defined by IMDb isn’t coming from an assigned authority; users submit genre tags based on guideline documentation, and these contributions are either approved or rejected by IMDb. Though this can feel like a limitation, understanding ‘genre’ as user-tagged metadata can add some nuance to analysis and potentially encourage future research of IMDb’s system.

I’ve opted to track ‘place of birth,’ which denotes where the director was born (via Wikipedia), instead of ‘nationality.’ To me, this avoids the thorny questions of citizenship.

**Release 📽️**

I’ve enriched this dataset with variables that may have interesting long-term value: genre, user\_review, and rating, which are all extracted from IMDb user-generated content. They reflect, in rough-and-ready shorthand, how a film is seen (or not seen) outside of the critical milieu that SaS’s poll purports to represent. A future researcher may want to revisit these data in 2025 or 2032 to see if SaS’s list has affected any shift in public perceptions or engagement. I hope this also serves as encouragment for other types of data collection that could further enhance this list and its potential in quantitative analysis.