Table W3.1: Data Availability Assessment

Issue		Feature	Reason for Importance Issue	Common Pitfalls and Complexities
1) Entity coverage and linkages				
1.1	For which entities is data available?	In Boxoffice: Rank, Title, Gross, Max Theater, Opening, Opening theathers, Opening thetater % of total gross, Distributor, Open. In IMDB: Title, Rating, Year	Incorporating a wider set of entities or entities that have not been used in prior research can lead to novel predictions and the exploration of new phenomena.	Data entities are available on different URLs which could make it complex to couple correct information and keep it neat and clear. Boxofficemojo is owned by IMDb and some data are quite limited (e.g. you have to do a subscription ad IMDb to access at all the data).
1.2	How many instances of an entity are available, and can they all be retrieved?	Boxofficemojo displays films on their website for every year. Users can also filter the films depends on the entity. IMDB displays all the film on their website for every year.	When no retrieval limits are present, it is easrier to create a sample and makes the research more reliable because of a bigger sample size.	Since the user has to scroll down and change page to be able to load and see the rest of the films, this could cause difficulty in scraping all entities wanted. So additional code will be needed to make sure data is scraped correctly.
1.3	How are instances of an entitiy identified (IDs)?	At boxofficemojo.com, films are indentified by their title. In the research bar users can only look up films by their name.	Consistent identification is essential when collecting data from the same website more than once.	Since IDs are unavailable in Boxofficemojo website, a consequent way of matching the correct entieties from Boxofficemojo to IMDB is required.
1.4	How are entities linked to one another?	Each film is linked to IMDb pro to see more details. For boxofficemojo every film is linked to cast information, crew information, company information, news and performance on IMDb pro website.	Likages are needed to make sure all data corresponds with the correct data on IMDp pro.	Improper linking of entities can limit the potential to create covariates from the raw data.
1.5	Can entities be linked to external entities?	Entities are linked from Boxofficemojo to IMDb Pro.	External identifiers allow the researcher to combine data from different sources into a unique combination of data, offering publication potential	Some linkages can only be established approximately.
1.6	Which lists or views could serve as a starting point for the data collection (“seeds”)?	The data scraped from Box Office Mojo contains all the data that is available in the table we use. We therefore do not use a sample for the Box Office Mojo data. The data scraped from IMDb is a sample that only contains the 1500 most popular movies released each year. As there are approximately 10.000 movies available on IMDb for each year, we decided to work with a subset.	Access to seeds can facilitate collecting the entire population or sampling from the relevant population	The sample may not be representative. It is crucial to consider to which population one can generalize, given the chosen seeds.
2) Time coverage				
2.1	For what period is data available?	The infnromation about the films which are shown on the website is real-time data.	It is important to be aware that data on the website might chance after scraping the website. The fact that it is real-time data, might improve the data and analysis.	Different time frames for different entities or variables. Obtaining historical data can be costly
2.2	How (accurate) is time encoded?	Boxofficemojo and IMDB timestamps are provided in PTD (Pacific Daylight Time).	It is important to understand the timestamp correctly, so that data can be compared and that the data is well interpreted.	Accurate encoding of time is necessary, especially when merging with other data. Some timestamps are inaccurately aggregated. Other timestamps may reflect not actual events but rather the time a data source started to exist. Inaccurate timestamps may require real-time scraping.
2.3	Can data be modified after it has been published?	Data can be modified after it has been published since different entities may change over time.	This means that the data is in real time as it is possible for the data to change at any time.	The analysis might not be the same after rating have been adjusted compared to the data that is scraped at a specific moment. So, if the website is scraped at different moments,the data could be different, and therefore the analysis could be too.
2.4	How often is the data refreshed?	Boxofficemojo is update daily. IMDB rating is update in real time.	Aggregation or interpolation are necessary since the data could change.	Being aware of refreshment timing is essential when deciding how to aggregate data temporally.
3) Algorithmic transparency and control				
3.1	Which mechanisms (e.g., algorithms, design choices) affect the display of data?	In boxofficemojo all the entities can affect the data display. Fox example, film are listed by opening theathers or by gross. In IMDB, Popularity, A-Z, User Rating, Number of Votes, US Box Office, Runtime, Year and Release Date can affect data display.	Algorithms might pose significant challenges to estimating causal effects or internal validity.	Popularity and User Rating cannot be taken as a "generalizable" metric. Since the values could change, the data display could change over time.
3.2	Is it clear how metrics have been calculated?	It is clear how metrics have been calculated.	It is crucial to find out how metrics are measured, so that the right conclussions are made	Metrics can be challenging to understand.
3.3	Can the researcher exert control over data display?	The researcher can control the data display by filter the values with a different entity.	By changing the default setting, it is possible to find a better sample for the analysis	Something that appears to be a systematic linkage between different entities could result from personalization algorithms.

Table W3.2: Evaluation of Research Fit and Resource Use

Issue		Feature	Reason for Importance Issue	Common Pitfalls and Complexities
1) Sampling and generalizability				
1.1	How can instances of entities be sampled from the site?	Many sampling methods can be used, however, we will decide only to work with a subset.	It is important to select the correct sampling method, as choosing the wrong one can induce bias in your model.	Arranging and selecting the random samples is more difficult and time consuming than simple random sampling.
1.2	What sample size is required to test the predictions?	We don't use a sample, but a subset. The subset contain the best 50 films by gross revenues per year (from 2015 to the present) from Boxofficemojo website.	It is important to consider the minimum sample size and recognize any technical constraints to satisfy statistical power requirements	It is difficult to really specify what sample size we need in order to satisfy these requirements.
1.3	What is the technically feasible sample size?	The technically feasible sample size is determined by the desired sampling frequency, the retrieval limit, number of calls for each instance of an entity, and the numbers of computers used.	It is important to calculate what is actually possible with your web scraper. By calculating this you can see whether you meet your requested sample size.	The effective sample size can severely vary, even with minor changes to any of the input parameters. One can solve for any of the other target parameters
2) Construct measurement				
2.1	Can the constructs of interest be measured with the available data?	Yes, the construct of interest can be measured with this data.	It is critical to ensure the raw data on website is a valid operationalizzation of a costruct. If it is unclear how a metrics is computed, it will be a hurdle to convice reviwers about validity.	Potential trade-offs may arise between automated versus human processing when constructing measures.
3) Data structure and preparation				
3.1	What data is required to answer the desired research question?	The data from Boxofficemojo required to answer the reserach question are titles and grosses.	The goal of the research influences the type of data required.	Since the target data set is at the daily level, when and how often does one need to visit the site to prepare such data? Does the website's updating frequency correspond with the level of aggregation?
3.2	How can the raw data be converted to a dataset that can be analyzed?	We obtain the raw data by scraping the website. Next, we will prepare the data and transform it, so that it will end up in a column of rows and columns. The data will be stored in Python. Then, we will process this data and create an output that someone else can use.	It is important to think about this, because not all analysis can be run without any transformation. Further, storing the data in a sensible way will make the research process more efficient.	Choosing the wrong type of data analysis is a pitfall, which can cost a lot of time. Furthermore, a pitfall is the coding. Many mistakes can be made there, which can influence the research.
4) Resource use				
4.1	What are the development costs?	No development costs are required.	Researchers may invest substantial research time in developing the extraction software.	Using well-documented packages/libraries can abate development efforts. Avoid exposure to legal risk by design. Seek counsel with the institutional legal team early on in the process or obtain external legal advice
4.2	How much will it cost to run the scraper?	No cost to run the scraper are planned.	Budgeting upfront is important to ensure an efficient research process.	It could take more time than what has been planned.
4.3	How costly is it to maintain the extraction software?	No cost to maintain the extraction software.	If it does cost something and you miss data, that can lead to a quality loss.	Erros can occur if you extract data for a longer period of time.
4.4	Opportunity costs	We haven't found any alternative soruces, however should exist.	Maybe alternative dataset exist.	Researchers need to decide between collecting novel data and using preexisting data.