Your final report, which should include:
Improved versions of your two previous intermediate assignments

- Data and Metadata Profile
- Repository Profile

Additional Information related to our course topics:

- A recommended data citation

A description of any considerations for long-term preservation. This might include
Whether the file formats are proprietary or might be in danger of going obsolete
Whether any specific software is needed to open any of the files
A statement about a copyright license that would be appropriate for this data set
A statement about any human subject considerations that apply to this data set. This might include:
Whether the data have personally identifiable data about people
Whether any steps were taken to anonymize or otherwise adjust the data for privacy or ethical reasons

LIS 545 B Term Project

Greeshma Elachitaya

Data and Metadata Profile

My Kaggle Dataset - US Film Industry Top Movies & Directors

URL: https://www.kaggle.com/datasets/thedevastator/us-film-industry-top-movies-directors

Data

What are the data?

The top films and directors in the US film industry between the years 2006 to 2017 are detailed in the dataset "US Film Industry - Top Movies & Directors" on Kaggle. By analyzing this comprehensive data, users can understand what types of films were popular over these years - from their budget sources to production formats. Furthermore, it gives an interesting look into trends in the filming industry over this time period.

What is the origin of the data?

The data was collected through two primary sources- **IMDb** and **The Numbers**.

Who are the key stakeholders?

The Kaggle account 'The Devastator' is associated with the dataset as an "owner" and "collaborator". The author of the original data has been credited as data.world's Admin (https://data.world/dataworldadmin).

How many data files are there? What do they contain? What file format are they?
There are **two** data files in the dataset.
movies.csv: This file contains information about the top 100 movies in the US film industry. It contains columns for the title, year of release, box office revenue, budget, and other details.
directors.csv: This file contains information about the directors of the top 100 movies in the US film industry. It contains columns for the director's name, nationality, and the number of movies they have directed.
Both the files are in the comma-separated values (**CSV**) file format, which can be read and processed by various software programs, including spreadsheets, databases, and programming languages like Python and R. Each row represents a single record, and each column represents a different field or attribute of that record.

Does the data set come with any usage restrictions? If so, what are they?
Per my research and understanding of the dataset, it has no usage restrictions. The owner has only mentioned the original author - data.world's Admin, be credited when the data is used. No license is required.

Is there any specific software required to open or analyze the data files?
Since it is a CSV file, it can be opened and analyzed using **Microsoft Excel, Google Sheets, LibreOffice Calc**, or any **other spreadsheet software**.

Metadata

What metadata does the data come with?
The data comes with metadata about the information about the structure and content of the data, i.e. the **names of columns** in both data files, along with a **short description** and **data type** (integer/string) of the values in each column.
It also includes the **collaborator** information (The Devastator), **provenance** metadata - which includes the source of the data (data.world) and **collection** methodology - scraping of data from data.world. Additionally, licensing metadata - all of the data has been made available under MIT **license** for anyone to use.
The two files also have metadata on the t**otal unique values, valid, missing or mismatched data** in each column, along with the **mean, standard deviation and quantiles** statistics for each column.

Where is the metadata?

The metadata has been provided in the dataset url on Kaggle under a separate section labelled metadata. The data files also include metadata on the column names and value descriptions.

How comprehensive is the metadata provided?
The metadata is quite comprehensive. The basic metadata has been provided, along with greater detail about the structure and content such as statistics on the data (mean, median, standard deviation).
But there could also have been more metadata on the history and context of the data.

Is the metadata structured according to any metadata standard?
The metadata is structured as per Kaggle norms for specifying metadata when creating new datasets and dataset versions, which is the **Data Package** specification ("Dataset Metadata · Kaggle/kaggle-api Wiki · GitHub") ("Metadata Standards Directory").

How could the current data and/or metadata be enriched? In particular, what additional information could be provided to:
improve users' ability to discover the data set in the repository environment.
assist somebody unfamiliar with the data to make use of the dataset for new purposes.

- **Keywords and subject terms**: Keywords and subjects that describe the data and its contents could be added into the metadata. This would increase users' ability to discover the dataset by linking it to similar datasets.
- **Data citation information**: The metadata could be further enriched by adding information on properly citing the data, such as date published, DOIs, etc.
- **Usage information**: Users unfamiliar with the data could be aided in making use of the dataset by adding usage information to the metadata. This could include information on how the dataset has been used in the past, including publications and studies that utilized the data.

What publications have been written (if any) based on this dataset?
No publications have been written based on this dataset.

Are any publications listed/provided with the data set?
No

What publications can you find using general web search engine that cite or otherwise reference the use of this dataset? How did you search for these publications (if any)?

Despite using Google Search, Bing, and DuckDuckGo; I was unable to find any citations or references to the use of this dataset.

References:

"Dataset Metadata · Kaggle/kaggle-api Wiki · GitHub." *GitHub*, 8 July 2021, https://github.com/Kaggle/kaggle-api/wiki/Dataset-Metadata.

"Metadata Standards Directory." *Research Data Alliance GitHub*, http://rd-alliance.github.io/metadata-directory/standards/.

"US Film Industry Top Movies & Directors." *Kaggle*, 22 January 2023, https://www.kaggle.com/datasets/thedevastator/us-film-industry-top-movies-directors.

Repository Profile

Repository - Open Science Framework (OSF)

URL: https://osf.io/

Repository's policies & procedures:

Why did you choose this repository?

1) The OSF is a free, open-source platform for researchers to manage, share, and archive their research data.
2) OSF supports a wide range of file formats and data types, which includes tabular data of the US Film Industry Top Movies and Directors data set.
3) OSF provides version control, which is essential for managing research data over time. This feature helps track changes and record these modifications to the dataset.
4) OSF assigns a unique and persistent identifier to each project, making citing and locating the data set easy.
5) It supports compliance with metadata standards, ensuring that the data set's metadata is consistent and standardized, making it easier for others to understand and reuse the dataset.
6) OSF allows for easy collaboration and sharing with researchers/public.

Is the repository open for data submissions from anybody, or does it have a defined collection scope?

The OSF is open for data submissions from anybody. It is a general-purpose research data repository for facilitating data sharing and collaboration across all disciplines. It doesnt have

any defined collection scope and has data from various fields. OSF allows for public or private sharing of data, depending on the user's preferences. (“Getting Started FAQ’s”)

Since the repository is open for submissions:

What data will the repository accept?

OSF accepts various research data from all fields and disciplines, including observational, experimental, compilation, and simulation data. There are no stated limits to what can be deposited on OSF, as long as the data does not violate any laws or ethical standards. There are some guidelines and requirements related to data types, domains, and file formats. (“Creating a data management plan (DMP) document”)

- Data types: OSF accepts many data types, including quantitative and qualitative data, software code, images, audio and video files, etc. However, sensitive or confidential data, such as personally identifiable information, medical records, or trade secrets, should not be deposited on OSF.
- Domains: OSF is designed to be a general-purpose research data repository, and it accepts data from all fields and disciplines, including natural sciences, social sciences, humanities, and engineering. But if the disciplines or domains have specific requirements or standards for data sharing or management, then users are required to follow them.
- File formats: OSF supports a wide range of file formats, including tabular data formats like CSV and Excel, text files, image files, audio and video files, etc. They recommend using open and non-proprietary file formats that are widely used and well-documented.

What guidance does the repository provide to a potential data submitter regarding what should be in the SIP?

- Basic metadata: The SIP should include basic metadata about the data. Data submitters can use existing standards of the discipline when possible. When there are no standards, describe the metadata that will be created.
- Documentation: The SIP should include documentation that describes the data, such as a data dictionary, codebook, or README file. This documentation should explain the meaning and structure of the data and any conventions or assumptions used in collecting or analyzing the data.The dataset should be categorized, described, and categorized in terms of stability of the dataset. They should be named uniquely and referenced.

- Code and software: If the data is associated with code or software, the SIP should include the code or software and documentation that describes how to use it.
- Licensing: The SIP should include information about the license or terms of use for the data.
- Preservation: The SIP should include information about the file formats, file naming conventions, and other technical details related to the long-term data preservation.
- Privacy and confidentiality: The SIP should include information about how privacy and confidentiality will be handled, including any direct/indirect identifiers, compliance with HIPAA, consent, etc.

(“Creating a data management plan (DMP) document”)

Does the repository provide any human assistance or consulting to the submitter?
OSF provides human assistance and consulting to users who need help depositing or managing their data on the platform. Users can contact the OSF support team via email, chat, or the OSF help center.

Does the repository require metadata to be submitted in any specific structure or according to any specific standard?
OSF utilizes a unique metadata model that facilitates FAIRness (Findable, Accessible, Interoperable, Reusable) and enables connections across the research lifecycle. The OSF Metadata Profile describes the community vocabularies and persistent identifiers that the OSF utilizes, the relationships available between metadata fields, the metatags used to enable enhanced web discovery, and an overall map of the metadata implementation. Data submitters can use existing standards of the discipline when possible. When there are no standards, they are required to describe the metadata that will be created.

Examine the repository's data access mechanisms

Is a login required to download data? If so, what is the process?
Depending on the access settings chosen by the data owner, users may or may not need to create a login to download data from OSF. If the data is publicly accessible, users can download the data without creating an account or logging in to the OSF platform. If the data owner has restricted access to the data, users must create an OSF account and be granted permission to access the data. The process is as follows:

1) Go to the OSF website (https://osf.io/) and click "Sign Up"
2) Choose a sign-up option

3) Follow the prompts to complete the sign-up process
4) Once the account is created, access to the restricted data can be requested by contacting the data owner.

Is more than one access mechanism provided?
Multiple access mechanisms are provided like direct file download, APIs and automated scripts, database query, and third party tools and plugins. The availability of the options depends on the access provided by the data owner.

Does the repository display metadata using any specific metadata standard?
OSF uses multiple metadata standards to display metadata, including Dublin Core, DataCite, Schema.org, etc.

What is included in this repository's DIP?
I could not find a DIP package specific to OSF.

References:

OSF Support, https://help.osf.io/article/573-osf-metadata-profile.

"Creating a data management plan (DMP) document." *OSF Support*, 23 March 2022, https://help.osf.io/article/144-creating-a-data-management-plan-dmp-document.

"Getting Started FAQ's." *OSF Support*, 15 December 2022, https://help.osf.io/article/546-getting-started-faq-s.

Additional Information

Recommended data citation?
Below is an example citation from the dataset in the APA style.
TheDevastator. (2020). US Film Industry: Top Movies & Directors. Retrieved February 20, 2023, from Kaggle website:
https://www.kaggle.com/thedevastator/us-film-industry-top-movies-directors

A description of any considerations for long-term preservation. This might include:

- *Whether the file formats are proprietary or might be in danger of going obsolete*
- *Whether any specific software is needed to open any of the files*

Considerations for long-term preservation:
File formats: It is essential to ensure that the file formats used to store the data are open, widely supported, and not proprietary in order to ensure that the data can be accessed and used by anyone in the future. For this dataset, the file format is CSV, which is a widely supported format for tabular data.
Software requirements: It is crucial to preserve any software required to access and use the data to ensure that it can be accessed and used in the future. A software package such as Excel or R could be used to open and analyze the CSV data files from this dataset.

A statement about a copyright license that would be appropriate for this data set:
The MIT license is a permissive license that allows for the free use, modification, and distribution of the dataset, subject to certain conditions such as attribution and disclaimer of liability. This license is appropriate for this dataset since the dataset is publicly available for general use and reuse. It is essential for any user of the dataset to abide by the terms and conditions of the MIT license.

A statement about any human subject considerations that apply to this data set. This might include:

- *Whether the data have personally identifiable data about people*
- *Whether any steps were taken to anonymize or otherwise adjust the data for privacy or ethical reasons*

This dataset does not contain personally identifiable data about people; therefore, no steps need to be taken to anonymize or adjust the data for privacy or ethical reasons. However, this dataset contains information about individuals and entities (Movie directors, etc.), and any use of the data should be done per applicable laws and ethical considerations. If the dataset is used to make any inferences or conclusions about the individuals or entities mentioned in the dataset, users should be careful to ensure that the inferences or conclusions are accurate and do not cause harm or offense to any individuals part of the dataset.