# notebook_igor

January 26, 2026

[ ]:
```
# in case export to pdf doens't work
#!sudo apt-get install texlive-xetex texlive-fonts-recommended␣
 ↪texlive-plain-generic
```

[ ]:
```
import kagglehub
from kagglehub import KaggleDatasetAdapter
import pandas as pd
import ast
```

[7]:
```
def load_csv(file_path):
    return kagglehub.load_dataset(
        KaggleDatasetAdapter.PANDAS,
        "rounakbanik/the-movies-dataset",
        file_path,
    )

movies = load_csv("movies_metadata.csv")
credits = load_csv("credits.csv")
keywords = load_csv("keywords.csv")
links = load_csv("links.csv")
links_small = load_csv("links_small.csv")
```

```
/tmp/ipykernel_185671/1641172065.py:2: DeprecationWarning: Use dataset_load()
instead of load_dataset(). load_dataset() will be removed in a future version.
  return kagglehub.load_dataset(

Downloading to /home/igorg/.cache/kagglehub/datasets/rounakbanik/the-movies-
dataset/versions/7/movies_metadata.csv…

100%|      | 12.2M/12.2M [00:02<00:00, 5.59MB/s]

Extracting zip of movies_metadata.csv…


/home/igorg/.pyenv/versions/3.10.6/envs/floportop/lib/python3.10/site-
packages/kagglehub/pandas_datasets.py:92: DtypeWarning: Columns (10) have mixed
types. Specify dtype option on import or set low_memory=False.
  result = read_function(
/tmp/ipykernel_185671/1641172065.py:2: DeprecationWarning: Use dataset_load()
```

```
instead of load_dataset(). load_dataset() will be removed in a future version.
  return kagglehub.load_dataset(

Downloading to /home/igorg/.cache/kagglehub/datasets/rounakbanik/the-movies-
dataset/versions/7/credits.csv…

100%|      | 43.1M/43.1M [00:08<00:00, 5.63MB/s]

Extracting zip of credits.csv…


/tmp/ipykernel_185671/1641172065.py:2: DeprecationWarning: Use dataset_load()
instead of load_dataset(). load_dataset() will be removed in a future version.
  return kagglehub.load_dataset(

Downloading to /home/igorg/.cache/kagglehub/datasets/rounakbanik/the-movies-
dataset/versions/7/keywords.csv…

100%|      | 1.39M/1.39M [00:00<00:00, 3.18MB/s]

Extracting zip of keywords.csv…


/tmp/ipykernel_185671/1641172065.py:2: DeprecationWarning: Use dataset_load()
instead of load_dataset(). load_dataset() will be removed in a future version.
  return kagglehub.load_dataset(

Downloading to /home/igorg/.cache/kagglehub/datasets/rounakbanik/the-movies-
dataset/versions/7/links.csv…

100%|      | 966k/966k [00:00<00:00, 1.34MB/s]
/tmp/ipykernel_185671/1641172065.py:2: DeprecationWarning: Use dataset_load()
instead of load_dataset(). load_dataset() will be removed in a future version.
  return kagglehub.load_dataset(

Downloading to /home/igorg/.cache/kagglehub/datasets/rounakbanik/the-movies-
dataset/versions/7/links_small.csv…

100%|      | 179k/179k [00:00<00:00, 564kB/s]
```

[9]:
```
movies = movies[movies["id"].str.isnumeric()]
movies["id"] = movies["id"].astype(int)

credits["id"] = credits["id"].astype(int)
keywords["id"] = keywords["id"].astype(int)
```

[10]:
```
links["imdbId"] = links["imdbId"].apply(lambda x: f"tt{x:07d}")
links_small["imdbId"] = links_small["imdbId"].apply(lambda x: f"tt{x:07d}")
```

[12]:
```
df = (
    movies
    .merge(credits, on="id", how="left")
    .merge(keywords, on="id", how="left")
```

```
    .merge(links, left_on="id", right_on="tmdbId", how="left")
)
df.head()
```

[12]:

```
   adult                              belongs_to_collection    budget  \
0  False  {'id': 10194, 'name': 'Toy Story Collection', …  30000000
1  False                                                NaN  65000000
2  False  {'id': 119050, 'name': 'Grumpy Old Men Collect…         0
3  False                                                NaN  16000000
4  False  {'id': 96871, 'name': 'Father of the Bride Col…         0

                                              genres  \
0  [{'id': 16, 'name': 'Animation'}, {'id': 35, '…
1  [{'id': 12, 'name': 'Adventure'}, {'id': 14, '…
2  [{'id': 10749, 'name': 'Romance'}, {'id': 35, …
3  [{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam…
4                    [{'id': 35, 'name': 'Comedy'}]

                             homepage     id    imdb_id original_language  \
0  http://toystory.disney.com/toy-story    862  tt0114709                en
1                                   NaN   8844  tt0113497                en
2                                   NaN  15602  tt0113228                en
3                                   NaN  31357  tt0114885                en
4                                   NaN  11862  tt0113041                en

               original_title  \
0                   Toy Story
1                     Jumanji
2            Grumpier Old Men
3           Waiting to Exhale
4  Father of the Bride Part II

                                          overview  …  \
0  Led by Woody, Andy's toys live happily in his …  …
1  When siblings Judy and Peter discover an encha…  …
2  A family wedding reignites the ancient feud be…  …
3  Cheated on, mistreated and stepped on, the wom…  …
4  Just when George Banks has recovered from his …  …

                        title  video vote_average vote_count  \
0                   Toy Story  False          7.7     5415.0
1                     Jumanji  False          6.9     2413.0
2            Grumpier Old Men  False          6.5       92.0
3           Waiting to Exhale  False          6.1       34.0
4  Father of the Bride Part II  False          5.7      173.0

                                               cast  \
```

```
0  [{'cast_id': 14, 'character': 'Woody (voice)',…
1  [{'cast_id': 1, 'character': 'Alan Parrish', '…
2  [{'cast_id': 2, 'character': 'Max Goldman', 'c…
3  [{'cast_id': 1, 'character': "Savannah 'Vannah…
4  [{'cast_id': 1, 'character': 'George Banks', '…

                                                crew  \
0  [{'credit_id': '52fe4284c3a36847f8024f49', 'de…
1  [{'credit_id': '52fe44bfc3a36847f80a7cd1', 'de…
2  [{'credit_id': '52fe466a9251416c75077a89', 'de…
3  [{'credit_id': '52fe44779251416c91011acb', 'de…
4  [{'credit_id': '52fe44959251416c75039ed7', 'de…

                                        keywords movieId     imdbId  \
0  [{'id': 931, 'name': 'jealousy'}, {'id': 4290,…       1  tt0114709
1  [{'id': 10090, 'name': 'board game'}, {'id': 1…       2  tt0113497
2  [{'id': 1495, 'name': 'fishing'}, {'id': 12392…       3  tt0113228
3  [{'id': 818, 'name': 'based on novel'}, {'id':…       4  tt0114885
4  [{'id': 1009, 'name': 'baby'}, {'id': 1599, 'n…       5  tt0113041

    tmdbId
0    862.0
1   8844.0
2  15602.0
3  31357.0
4  11862.0

[5 rows x 30 columns]
```

[13]:
```
def parse_json(col):
    return col.apply(lambda x: ast.literal_eval(x) if pd.notna(x) else [])
```

[14]:
```
df["genres"] = parse_json(df["genres"])
df["keywords"] = parse_json(df["keywords"])
df["cast"] = parse_json(df["cast"])
df["crew"] = parse_json(df["crew"])
```

[15]:
```
df.sample()
```

[15]:
```
      adult belongs_to_collection    budget  \
2235  False                   NaN  14000000

                                              genres homepage    id  \
2235  [{'id': 18, 'name': 'Drama'}, {'id': 53, 'name…      NaN  9445

        imdb_id original_language original_title  \
2235  tt0118636                en      Apt Pupil
```

```
                                                 overview  …      title  \
2235  Neighborhood boy Todd Bowden discovers that an…  …  Apt Pupil

      video vote_average vote_count  \
2235  False          6.2      168.0

                                                   cast  \
2235  [{'cast_id': 10, 'character': 'Kurt Dussander'…

                                                   crew  \
2235  [{'credit_id': '52fe44f8c3a36847f80b4eab', 'de…

                                               keywords movieId     imdbId  \
2235  [{'id': 1308, 'name': 'secret identity'}, {'id…    2320  tt0118636

      tmdbId
2235  9445.0

[1 rows x 30 columns]
```

[16]:
```
df["genre_names"] = df["genres"].apply(
    lambda xs: [x["name"] for x in xs]
)
```

[18]:
```
df['genre_names'].sample(4)
```

[18]:
```
16933    [Horror, Mystery, Thriller]
24236                        [Horror]
28545                         [Drama]
4928                          [Drama]
Name: genre_names, dtype: object
```

[19]:
```
df["keyword_names"] = df["keywords"].apply(
    lambda xs: [x["name"] for x in xs]
)
```

[20]:
```
df['keyword_names'].sample(4)
```

[20]:
```
36501    [fight, cop, love, thief, revenge, look-alike,…
20080                                                 []
4068     [casino, submachine gun, hold-up robbery, elvi…
37578                                                 []
Name: keyword_names, dtype: object
```

[21]:
```
df["cast_top"] = df["cast"].apply(
    lambda xs: [x["name"] for x in xs[:10]]
```

```
)
```

[22]:
```
df['cast_top'].sample(4)
```

[22]:
```
585      [Jack Nicholson, Michael Keaton, Kim Basinger,…
8913     [David Bowie, Tom Conti, Ryuichi Sakamoto, Tak…
31742    [Katherine Heigl, Ben Barnes, Clea DuVall, She…
9180     [Rick Gianasi, Susan Byun, Bill Weeden, Thomas…
Name: cast_top, dtype: object
```

[23]:
```
df["directors"] = df["crew"].apply(
    lambda xs: [x["name"] for x in xs if x["job"] == "Director"]
)
```

[24]:
```
df["year"] = pd.to_datetime(
    df["release_date"], errors="coerce"
).dt.year

df["runtime"] = pd.to_numeric(df["runtime"], errors="coerce")
df["vote_average"] = pd.to_numeric(df["vote_average"], errors="coerce")
df["vote_count"] = pd.to_numeric(df["vote_count"], errors="coerce")
df["popularity"] = pd.to_numeric(df["popularity"], errors="coerce")
```

[25]:
```
wide_df = df[[
    "id",                 # TMDB ID
    "imdbId",             # IMDb ID (primary external key)
    "title",
    "original_title",
    "overview",
    "genre_names",
    "keyword_names",
    "cast_top",
    "directors",
    "original_language",
    "year",
    "runtime",
    "vote_average",
    "vote_count",
    "popularity",
]].reset_index(drop=True)
```

[26]:
```
wide_df.sample(3)
```

[26]:
```
          id     imdbId                         title  \
28742  24633  tt1123970                  Phantom Pain
21149  77221  tt1701210                    Black Gold
2570   36685  tt0073629  The Rocky Horror Picture Show
```

```
                    original_title  \
28742                Phantomschmerz
21149                    Black Gold
2570   The Rocky Horror Picture Show

                                                 overview  \
28742  Marc is a passionate cyclist and urban slacker…
21149  On the Arabian Peninsula in the 1930s, two war…
2570   Sweethearts Brad and Janet, stuck with a flat …

                                genre_names  \
28742                        [Drama, Foreign]
21149                      [Adventure, Drama]
2570   [Comedy, Horror, Music, Science Fiction]

                                      keyword_names  \
28742                                            []
21149                                            []
2570   [transvestism, transylvania, sex, marriage pro…

                                                 cast_top  \
28742  [Til Schweiger, Jana Pallaske, Stipe Erceg, Ju…
21149  [Mark Strong, Antonio Banderas, Freida Pinto, …
2570   [Tim Curry, Susan Sarandon, Barry Bostwick, Ri…

                 directors original_language    year  runtime  vote_average  \
28742      [Matthias Emcke]                de  2009.0     97.0           5.3
21149  [Jean-Jacques Annaud]                en  2011.0    130.0           5.9
2570           [Jim Sharman]                en  1975.0    100.0           7.4

       vote_count  popularity
28742         8.0    1.161097
21149        77.0    6.475665
2570        703.0    8.699428
```

[28]:
```
wide_df["embedding_text"] = (
    "Title: " + wide_df["title"].fillna("") + "\n"
    "Overview: " + wide_df["overview"].fillna("") + "\n"
    "Genres: " + wide_df["genre_names"].apply(lambda x: ", ".join(x)) + "\n"
    "Keywords: " + wide_df["keyword_names"].apply(lambda x: ", ".join(x)) + "\n"
    "Cast: " + wide_df["cast_top"].apply(lambda x: ", ".join(x)) + "\n"
    "Director: " + wide_df["directors"].apply(lambda x: ", ".join(x))
)
```

[29]:
```
wide_df["embedding_text"].sample(4)
```

[29]:
```
15777    Title: A Name for Evil\nOverview: Dissatisfied…
15695    Title: Kurt Cobain: About a Son\nOverview: An …
45951    Title: After Love\nOverview: After 15 years of…
42275    Title: Moondance Alexander\nOverview: The curi…
Name: embedding_text, dtype: object
```

[ ]: