eda_pdf

August 4, 2025

[1]:
```
# Cell 1: Setup and Imports
import os
import logging
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from scipy.stats import shapiro
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, count, when, split, explode, year, log,␣
 ↪coalesce, to_date, lit, concat, rand, mean
from pyspark.sql.types import StructType, StructField, IntegerType, StringType,␣
 ↪FloatType
```

[2]:
```
# Configure logging
logging.basicConfig(
    filename='eda_pipeline.log',
    level=logging.INFO,
    format='%(asctime)s - %(levelname)s - %(message)s',
    datefmt='%Y-%m-%d %H:%M:%S'
)
console = logging.StreamHandler()
console.setLevel(logging.INFO)
console.setFormatter(logging.Formatter(
    '%(asctime)s - %(levelname)s - %(message)s'))
logging.getLogger().addHandler(console)

# Stop any existing Spark sessions
spark = SparkSession.getActiveSession()
if spark:
    spark.stop()
    logging.info("Stopped existing Spark session")

# Initialize Spark session
os.environ["PYSPARK_PYTHON"] = os.path.abspath(".venv/Scripts/python.exe")

spark = SparkSession.builder \
```

```
    .appName("CostAwareRecommendation") \
    .master("local[*]") \
    .config("spark.driver.memory", "8g") \
    .config("spark.pyspark.python", os.environ["PYSPARK_PYTHON"]) \
    .config("spark.driver.bindAddress", "127.0.0.1") \
    .config("spark.driver.host", "127.0.0.1") \
    .config("spark.driver.port", "25334") \
    .config("spark.ui.showConsoleProgress", "false") \
    .config("spark.driver.extraJavaOptions", "-Xss4m") \
    .config("spark.executor.extraJavaOptions", "-Xss4m") \
    .getOrCreate()
logging.info(f"Created Spark session: {spark.sparkContext.uiWebUrl}")

# Set plotting style
sns.set_style("whitegrid")
plt.rcParams['figure.figsize'] = (10, 6)

logging.info(
    "Spark session initialized, logging configured, and plotting style set.")
```

```
2025-08-03 19:17:05,030 - INFO - Created Spark session: http://127.0.0.1:4041
2025-08-03 19:17:05,030 - INFO - Spark session initialized, logging configured,
and plotting style set.
```

[3]:
```
def load_data_and_budgets():
    try:
        # Define schemas
        ratings_schema = StructType([
            StructField("userId", IntegerType(), False),
            StructField("movieId", IntegerType(), False),
            StructField("rating", FloatType(), False),
            StructField("timestamp", IntegerType(), False)
        ])
        movies_schema = StructType([
            StructField("movieId", IntegerType(), False),
            StructField("title", StringType(), False),
            StructField("genres", StringType(), True)
        ])
        links_schema = StructType([
            StructField("movieId", IntegerType(), False),
            StructField("imdbId", StringType(), True),
            StructField("tmdbId", StringType(), True)
        ])
        title_basics_schema = StructType([
            StructField("tconst", StringType(), False),
            StructField("titleType", StringType(), True),
            StructField("primaryTitle", StringType(), True),
```

```
            StructField("originalTitle", StringType(), True),
            StructField("isAdult", StringType(), True),
            StructField("startYear", StringType(), True),
            StructField("endYear", StringType(), True),
            StructField("runtimeMinutes", StringType(), True),
            StructField("genres", StringType(), True)
        ])

        # Load MovieLens data
        ratings_df = spark.read.schema(ratings_schema).csv(
            "ml-25m/ratings.csv", header=True)
        movies_df = spark.read.schema(movies_schema).csv(
            "ml-25m/movies.csv", header=True)
        links_df = spark.read.schema(links_schema).csv(
            "ml-25m/links.csv", header=True)

        # Load IMDb data
        basics_df = spark.read.schema(title_basics_schema).option(
            "delimiter", "\t").csv("imdb/title.basics.tsv.gz", header=True)

        # Check if budgets CSV exists
        budgets_path = os.path.abspath("tmdb_budgets.csv")

        budgets_df = spark.read.csv(
            budgets_path, header=True, inferSchema=True)

        # Format imdbId to match tconst (add 'tt' prefix)
        links_df = links_df.withColumn(
            "imdbId", concat(lit("tt"), col("imdbId")))

        logging.info("Loaded all datasets and budgets.")
        return ratings_df, movies_df, links_df, basics_df, budgets_df
    except Exception as e:
        logging.error(f"Failed to load data and budgets: {str(e)}")
        raise

ratings_df, movies_df, links_df, basics_df, budgets_df = load_data_and_budgets()
```

```
2025-08-03 19:17:13,481 - INFO - Loaded all datasets and budgets.
```

[12]:
```
print([df.show(10) for df in [ratings_df, movies_df, links_df, basics_df,␣
↪budgets_df]])
```

```
+------+-------+------+----------+
|userId|movieId|rating| timestamp|
+------+-------+------+----------+
|     1|      1|   4.0|1225734739|
|     1|    110|   4.0|1225865086|
```

```
|     1|    158|   4.0|1225733503|
|     1|    260|   4.5|1225735204|
|     1|    356|   5.0|1225735119|
|     1|    381|   3.5|1225734105|
|     1|    596|   4.0|1225733524|
|     1|   1036|   5.0|1225735626|
|     1|   1049|   3.0|1225734079|
|     1|   1066|   4.0|1225736961|
+------+-------+------+----------+
only showing top 10 rows
+-------+--------------------+--------------------+---------+------------+
|movieId|               title|              genres|   budget|release_date|
+-------+--------------------+--------------------+---------+------------+
| 167420|  Master i Margarita|       Drama,Fantasy|      0.0|  1994-06-06|
|   3751|         Chicken Run|Adventure,Animati…|    4.5E7|  2000-06-23|
|   5791|               Frida|Biography,Drama,R…|    1.2E7|  2002-08-29|
|  33493|Star Wars: Episod…|Action,Adventure,…|   1.13E8|  2005-05-17|
|   3457|     Waking the Dead|Drama,Mystery,Rom…|8500000.0|  2000-03-24|
|   4265|              Driven|  Action,Drama,Sport|    9.4E7|  2001-04-27|
|   4643|  Planet of the Apes|Action,Adventure,…|    1.0E8|  2001-07-25|
|   5882|     Treasure Planet|Adventure,Animati…|    1.4E8|  2002-11-26|
|   3273|            Scream 3|      Horror,Mystery|    4.0E7|  2000-02-04|
|   3300|         Pitch Black|Horror,Sci-Fi,Thr…|    2.3E7|  2000-02-18|
+-------+--------------------+--------------------+---------+------------+
only showing top 10 rows
+-------+---------+------+
|movieId|   imdbId|tmdbId|
+-------+---------+------+
|      1|tt0114709|   862|
|      2|tt0113497|  8844|
|      3|tt0113228| 15602|
|      4|tt0114885| 31357|
|      5|tt0113041| 11862|
|      6|tt0113277|   949|
|      7|tt0114319| 11860|
|      8|tt0112302| 45325|
|      9|tt0114576|  9091|
|     10|tt0113189|   710|
+-------+---------+------+
only showing top 10 rows
+---------+---------+--------------------+--------------------+-------+---------
+-------+-------------+--------------------+
|   tconst|titleType|        primaryTitle|
originalTitle|isAdult|startYear|endYear|runtimeMinutes|              genres|
+---------+---------+--------------------+--------------------+-------+---------
+-------+-------------+--------------------+
|tt0000001|    short|           Carmencita|           Carmencita|      0|
1894|     \N|             1|   Documentary,Short|
```

```
|tt0000002|    short|Le clown et ses c…|Le clown et ses c…|      0|
1892|     \N|             5|     Animation,Short|
|tt0000003|    short|        Poor Pierrot|      Pauvre Pierrot|      0|
1892|     \N|             5|Animation,Comedy,…|
|tt0000004|    short|          Un bon bock|          Un bon bock|      0|
1892|     \N|            12|     Animation,Short|
|tt0000005|    short|    Blacksmith Scene|    Blacksmith Scene|      0|
1893|     \N|             1|               Short|
|tt0000006|    short|   Chinese Opium Den|   Chinese Opium Den|      0|
1894|     \N|             1|               Short|
|tt0000007|    short|Corbett and Court…|Corbett and Court…|      0|
1894|     \N|             1|         Short,Sport|
|tt0000008|    short|Edison Kinetoscop…|Edison Kinetoscop…|      0|
1894|     \N|             1|   Documentary,Short|
|tt0000009|    movie|           Miss Jerry|           Miss Jerry|      0|
1894|     \N|            45|             Romance|
|tt0000010|    short| Leaving the Factory|La sortie de l'us…|      0|
1895|     \N|             1|   Documentary,Short|
+---------+---------+--------------------+--------------------+-------+---------
+-------+-------------+--------------------+
only showing top 10 rows
+-------+------+--------+------------+
|movieId|tmdbId|  budget|release_date|
+-------+------+--------+------------+
|      1|   862|30000000|  1995-11-22|
|      2|  8844|65000000|  1995-12-15|
|      3| 15602|25000000|  1995-12-22|
|      4| 31357|16000000|  1995-12-22|
|      5| 11862|       0|  1995-12-08|
|      6|   949|60000000|  1995-12-15|
|      7| 11860|58000000|  1995-12-15|
|      8| 45325|       0|  1995-12-22|
|      9|  9091|35000000|  1995-10-27|
|     10|   710|60000000|  1995-11-16|
+-------+------+--------+------------+
only showing top 10 rows
```

[12]:
```
[None, None, None, None, None]
```

[4]:
```python
import time

def preprocess_data(ratings_df, movies_df, links_df, basics_df, budgets_df):
    try:
        # Log start of preprocessing
        logging.info(f"Starting preprocess_data at {time.time()}")

        # Validate input DataFrames
```

```
        logging.info("Validating input DataFrames...")
        for df, name in [(ratings_df, "ratings_df"), (movies_df, "movies_df"),
                         (links_df, "links_df"), (basics_df, "basics_df"),
                         (budgets_df, "budgets_df")]:
            if df is None:
                raise ValueError(f"{name} is None")
            df_count = df.count()
            logging.info(f"{name} has {df_count} rows")

        # Clean startYear by replacing '\N' with null in basics_df
        basics_df = basics_df.withColumn(
            "startYear",
            when(col("startYear") != "\\N", col("startYear")).otherwise(None)
        )

        # Merge datasets
        movies_df = (
            movies_df
            .join(links_df, "movieId", "inner")
            .join(
                basics_df.select(
                    col("tconst"),
                    col("primaryTitle"),
                    col("startYear"),
                    col("genres").alias("basics_genres")
                ),
                col("imdbId") == col("tconst"),
                "left"
            )
            .join(budgets_df, "movieId", "left")
            .filter(
                (col("startYear").isNotNull() & col("startYear").cast("int").
↪between(2000, 2025)) |
                (col("release_date").isNotNull() & to_date(
                    col("release_date")).cast("string").like("20[0-2][0-5]%"))
            )
            .select(
                col("movieId"),
                col("primaryTitle").alias("title"),
                coalesce(col("basics_genres"), col("genres")).alias("genres"),
                col("budget").cast("float"),
                to_date(col("release_date")).alias("release_date")
            )
        )

        logging.info(
```

```
            f"Preprocessed {ratings_df.count()} ratings and {movies_df.count()}␣
 ↪movies.")
        return ratings_df, movies_df
    except Exception as e:
        logging.error(f"Failed to preprocess data: {str(e)}")
        raise


# Call the function
ratings_df, movies_df = preprocess_data(
    ratings_df, movies_df, links_df, basics_df, budgets_df)
```

```
2025-08-03 19:17:14,236 - INFO - Starting preprocess_data at 1754266634.2368038
2025-08-03 19:17:14,243 - INFO - Validating input DataFrames…
2025-08-03 19:17:17,531 - INFO - ratings_df has 33832162 rows
2025-08-03 19:17:17,812 - INFO - movies_df has 86537 rows
2025-08-03 19:17:18,058 - INFO - links_df has 86537 rows
2025-08-03 19:17:26,962 - INFO - basics_df has 11771649 rows
2025-08-03 19:17:27,279 - INFO - budgets_df has 85305 rows
2025-08-03 19:18:13,411 - INFO - Preprocessed 33832162 ratings and 50888 movies.
```

[15]:
```
print("ratings_df", ratings_df.show(10))
print("movies_df", movies_df.show(10))
```

```
+------+-------+------+----------+
|userId|movieId|rating| timestamp|
+------+-------+------+----------+
|     1|      1|   4.0|1225734739|
|     1|    110|   4.0|1225865086|
|     1|    158|   4.0|1225733503|
|     1|    260|   4.5|1225735204|
|     1|    356|   5.0|1225735119|
|     1|    381|   3.5|1225734105|
|     1|    596|   4.0|1225733524|
|     1|   1036|   5.0|1225735626|
|     1|   1049|   3.0|1225734079|
|     1|   1066|   4.0|1225736961|
+------+-------+------+----------+
only showing top 10 rows
ratings_df None
+-------+--------------------+--------------------+---------+------------+
|movieId|               title|              genres|   budget|release_date|
+-------+--------------------+--------------------+---------+------------+
| 167420|  Master i Margarita|       Drama,Fantasy|      0.0|  1994-06-06|
|   3751|         Chicken Run|Adventure,Animati…|    4.5E7|  2000-06-23|
|   5791|               Frida|Biography,Drama,R…|    1.2E7|  2002-08-29|
|  33493|Star Wars: Episod…|Action,Adventure,…|   1.13E8|  2005-05-17|
|   3457|     Waking the Dead|Drama,Mystery,Rom…|8500000.0|  2000-03-24|
```

```
|   4265|              Driven|  Action,Drama,Sport|    9.4E7|  2001-04-27|
|   4643|  Planet of the Apes|Action,Adventure,…|    1.0E8|  2001-07-25|
|   5882|     Treasure Planet|Adventure,Animati…|    1.4E8|  2002-11-26|
|   3273|            Scream 3|      Horror,Mystery|    4.0E7|  2000-02-04|
|   3300|         Pitch Black|Horror,Sci-Fi,Thr…|    2.3E7|  2000-02-18|
+-------+--------------------+--------------------+---------+------------+
only showing top 10 rows
movies_df None
```

[9]:
```python
# Cell 3: Column Descriptions
def describe_columns():
    try:
        descriptions = {
            "ratings_df": {
                "userId": "Unique identifier for a user (integer).",
                "movieId": "Unique identifier for a movie (integer), links to␣
 ↪movies_df.",
                "rating": "User rating for the movie (float, 0.5 to 5.0, in 0.5␣
 ↪increments).",
                "timestamp": "Unix timestamp of when the rating was given␣
 ↪(integer, seconds since 1970-01-01)."
            },
            "merged_df": {
                "movieId": "Unique identifier for a movie (integer), links to␣
 ↪ratings_df.",
                "title": "Movie title (string, sourced from IMDb primaryTitle).
 ↪",
                "genres": "Pipe-separated list of genres (string, e.g.,␣
 ↪'Action|Drama').",
                "budget": "Production budget in USD (float, from TMDb, may be␣
 ↪missing).",
                "release_year": "Year of movie release (integer, 2000-2020,␣
 ↪from TMDb release_date)."
            }
        }

        # Log descriptions
        for df_name, cols in descriptions.items():
            logging.info(f"Column descriptions for {df_name}:")
            for col_name, desc in cols.items():
                logging.info(f"  {col_name}: {desc}")

        # Save to file for dissertation
        with open("eda_plots/column_descriptions.txt", "w") as f:
            for df_name, cols in descriptions.items():
                f.write(f"{df_name}:\n")
                for col_name, desc in cols.items():
```

```
                    f.write(f"  {col_name}: {desc}\n")

        logging.info(
            "Column descriptions saved to eda_plots/column_descriptions.txt")
    except Exception as e:
        logging.error(f"Failed to describe columns: {str(e)}")
        raise


describe_columns()
```

[13]:
```
# Cell 4: Data Size
def analyze_data_size(ratings_df, movies_df):
    try:
        # Count rows and columns
        ratings_rows, ratings_cols = ratings_df.count(), len(ratings_df.columns)
        movies_rows, movies_cols = movies_df.count(), len(movies_df.columns)

        # Estimate memory usage (approximate)
        ratings_size_mb = (ratings_rows * ratings_cols * 8) / \
            (1024 ** 2)  # Assuming 8 bytes per cell
        movies_size_mb = (movies_rows * movies_cols * 8) / (1024 ** 2)

        # Log results
        logging.info(
            f"ratings_df: {ratings_rows} rows, {ratings_cols} columns,␣
 ↪~{ratings_size_mb:.2f} MB")
        logging.info(
            f"movies_df: {movies_rows} rows, {movies_cols} columns,␣
 ↪~{movies_size_mb:.2f} MB")

        # Save to file
        with open("eda_plots/data_size.txt", "w") as f:
            f.write(
                f"ratings_df: {ratings_rows} rows, {ratings_cols} columns,␣
 ↪~{ratings_size_mb:.2f} MB\n")
            f.write(
                f"movies_df: {movies_rows} rows, {movies_cols} columns,␣
 ↪~{movies_size_mb:.2f} MB\n")

        logging.info("Data size saved to eda_plots/data_size.txt")
    except Exception as e:
        logging.error(f"Failed to analyze data size: {str(e)}")
        raise


analyze_data_size(ratings_df, movies_df)
```

```
2025-08-03 17:13:05,110 - INFO - ratings_df: 33832162 rows, 4 columns, ~1032.48
MB
2025-08-03 17:13:05,110 - INFO - movies_df: 50888 rows, 5 columns, ~1.94 MB
2025-08-03 17:13:05,119 - INFO - Data size saved to eda_plots/data_size.txt
```

[17]:
```python
# Cell 5: Distributions and Normality Tests
def analyze_distributions(ratings_df, movies_df):
    try:
        # Convert to Pandas for plotting and normality tests
        ratings_pd = ratings_df.select("rating").toPandas()
        movies_pd = movies_df.select(
            col("budget"),
            year("release_date").alias("release_year")
        ).toPandas()

        # Summary statistics
        ratings_stats = ratings_df.select("rating").describe().toPandas()
        budget_stats = movies_df.filter(col("budget") > 0).select(
            "budget").describe().toPandas()
        year_stats = movies_df.select(year("release_date").alias("release␣
 ↪year")).describe().toPandas()

        # Plot histograms
        plt.figure()
        sns.histplot(ratings_pd["rating"], bins=10, kde=True)
        plt.title("Distribution of Ratings")
        plt.xlabel("Rating")
        plt.ylabel("Count")
        plt.savefig("eda_plots/ratings_histogram.png")
        plt.close()

        plt.figure()
        sns.histplot(movies_pd[movies_pd["budget"] > 0]
                     ["budget"], bins=50, kde=True)
        plt.title("Distribution of Non-Zero Budgets")
        plt.xlabel("Budget (USD)")
        plt.ylabel("Count")
        plt.savefig("eda_plots/budget_histogram.png")
        plt.close()

        plt.figure()
        sns.histplot(movies_pd["release_year"], bins=21, kde=True)
        plt.title("Distribution of Release Years (2000-2020)")
        plt.xlabel("Release Year")
        plt.ylabel("Count")
        plt.savefig("eda_plots/release_year_histogram.png")
        plt.close()
```

```
        # Normality tests (Shapiro-Wilk on sample due to large data)
        rating_sample = ratings_pd["rating"].sample(
            n=min(5000, len(ratings_pd)), random_state=42)
        budget_sample = movies_pd[movies_pd["budget"] > 0]["budget"].sample(
            n=min(5000, len(movies_pd[movies_pd["budget"] > 0])),␣
↪random_state=42)
        year_sample = movies_pd["release_year"].dropna().sample(
            n=min(5000, len(movies_pd)), random_state=42)

        rating_stat, rating_p = shapiro(rating_sample)
        # Log-transform budget for normality test
        budget_stat, budget_p = shapiro(np.log1p(budget_sample))
        year_stat, year_p = shapiro(year_sample)

        # Log results
        logging.info("Rating distribution stats:\n" +
                     ratings_stats.to_string())
        logging.info(
            f"Rating normality test: W={rating_stat:.4f}, p={rating_p:.4f}␣
↪{'(not normal)' if rating_p < 0.05 else '(normal)'}")
        logging.info("Budget distribution stats:\n" + budget_stats.to_string())
        logging.info(
            f"Budget (log-transformed) normality test: W={budget_stat:.4f},␣
↪p={budget_p:.4f} {'(not normal)' if budget_p < 0.05 else '(normal)'}")
        logging.info("Release year distribution stats:\n" +
                     year_stats.to_string())
        logging.info(
            f"Release year normality test: W={year_stat:.4f}, p={year_p:.4f}␣
↪{'(not normal)' if year_p < 0.05 else '(normal)'}")

        # Save stats
        with open("eda_plots/distribution_stats.txt", "w") as f:
            f.write("Rating distribution stats:\n" +
                    ratings_stats.to_string() + "\n\n")
            f.write(
                f"Rating normality test: W={rating_stat:.4f}, p={rating_p:.4f}␣
↪{'(not normal)' if rating_p < 0.05 else '(normal)'}\n")
            f.write("Budget distribution stats:\n" +
                    budget_stats.to_string() + "\n\n")
            f.write(
                f"Budget (log-transformed) normality test: W={budget_stat:.4f},␣
↪p={budget_p:.4f} {'(not normal)' if budget_p < 0.05 else '(normal)'}\n")
            f.write("Release year distribution stats:\n" +
                    year_stats.to_string() + "\n\n")
            f.write(
```

```
                f"Release year normality test: W={year_stat:.4f}, p={year_p:.
↪4f} {'(not normal)' if year_p < 0.05 else '(normal)'}\n")

        logging.info(
            "Distributions and normality tests saved to eda_plots/
↪distribution_stats.txt")
    except Exception as e:
        logging.error(f"Failed to analyze distributions: {str(e)}")
        raise


analyze_distributions(ratings_df, movies_df)
```

```
2025-08-03 17:49:27,563 - INFO - Rating distribution stats:
  summary              rating
0   count            33832162
1    mean    3.54254040873888
2  stddev  1.0639586178664844
3     min                 0.5
4     max                 5.0
2025-08-03 17:49:27,568 - INFO - Rating normality test: W=0.9294, p=0.0000 (not
normal)
2025-08-03 17:49:27,568 - INFO - Budget distribution stats:
  summary                budget
0   count                 10367
1    mean  2.0266758294781517E7
2  stddev    3.73178623947418E7
3     min                   1.0
4     max               4.654E8
2025-08-03 17:49:27,578 - INFO - Budget (log-transformed) normality test:
W=0.8670, p=0.0000 (not normal)
2025-08-03 17:49:27,582 - INFO - Release year distribution stats:
  summary       release year
0   count              50017
1    mean  2012.900493832097
2  stddev  6.124272447228714
3     min               1951
4     max               2025
2025-08-03 17:49:27,582 - INFO - Release year normality test: W=0.9598, p=0.0000
(not normal)
2025-08-03 17:49:27,589 - INFO - Distributions and normality tests saved to
eda_plots/distribution_stats.txt
```

[19]:
```
# Cell 6: Sample Data
def sample_data(ratings_df, movies_df):
    try:
        # Collect 5 random samples
```

```
        ratings_sample = ratings_df.orderBy(rand()).limit(5).toPandas()
        movies_sample = movies_df.orderBy(rand()).limit(5).toPandas()

        # Log samples
        logging.info("Ratings sample:\n" + ratings_sample.to_string())
        logging.info("Movies sample:\n" + movies_sample.to_string())

        # Save to file
        with open("eda_plots/data_samples.txt", "w") as f:
            f.write("Ratings sample:\n" + ratings_sample.to_string() + "\n\n")
            f.write("Movies sample:\n" + movies_sample.to_string() + "\n")

        logging.info("Data samples saved to eda_plots/data_samples.txt")
    except Exception as e:
        logging.error(f"Failed to sample data: {str(e)}")
        raise


sample_data(ratings_df, movies_df)
```

```
2025-08-03 17:53:45,835 - INFO - Ratings sample:
   userId  movieId  rating   timestamp
0  277636   177765     4.0  1685470107
1  226948        5     4.0   982427965
2  239415    33834     3.5  1633713947
3  145639     1474     4.0   884338404
4   27069     4963     3.5  1094007447
2025-08-03 17:53:45,843 - INFO - Movies sample:
   movieId                                      title
genres      budget release_date
0     5402  Nijinsky: The Diaries of Vaslav Nijinsky
Biography,Documentary,Music         0.0   2002-04-25
1   277158                                 Alex's War
Biography,Documentary,History         0.0   2022-07-29
2   260855                               Naked Poison
Crime,Thriller         0.0   2000-09-23
3    94352                               Bride Flight
Drama,Romance   6400000.0   2008-10-15
4   185001                             I Feel Pretty
Comedy,Romance  32000000.0   2018-04-19
2025-08-03 17:53:45,849 - INFO - Data samples saved to
eda_plots/data_samples.txt
```

[6]:
```
# Cell 7: Additional EDA
def additional_eda(ratings_df, merged_df):
    try:
        # Missing value analysis
```

```
ratings_missing = ratings_df.select(
    [count(when(col(c).isNull(), c)).alias(c)
     for c in ratings_df.columns]
).toPandas()

movies_missing = merged_df.select(
    [count(when(col(c).isNull(), c)).alias(c)
     for c in merged_df.columns]
).toPandas()

# Unique value counts
unique_users = ratings_df.select("userId").distinct().count()
unique_movies_ratings = ratings_df.select("movieId").distinct().count()
unique_movies = merged_df.select("movieId").distinct().count()

# Genre distribution
genre_df = merged_df.select(
    explode(split("genres", "\\|")).alias("genre")
).groupBy("genre").count().toPandas()

# Correlation: Budget vs. average rating
avg_rating_df = ratings_df.groupBy("movieId").agg(
    mean("rating").alias("avg_rating")
)

corr_df = merged_df.join(avg_rating_df, "movieId", "inner").select(
    "budget", "avg_rating"
).filter(col("budget") > 0).toPandas()

correlation = corr_df["budget"].corr(corr_df["avg_rating"])

# Temporal trends: Ratings and budgets by year
ratings_by_year = ratings_df.join(
    merged_df.select("movieId", year(
        to_date("release_date")).alias("release_year")),
    "movieId"
).groupBy("release_year").agg(
    count("rating").alias("rating_count"),
    mean("rating").alias("avg_rating")
).toPandas()

# Budgets by year (for non-zero budgets)
budgets_by_year = merged_df.filter(col("budget") > 0).select(
    year(to_date("release_date")).alias("release_year"),
    "budget"
).groupBy("release_year").agg(
    mean("budget").alias("avg_budget")
```

```
        ).toPandas()

        # Plot genre distribution
        plt.figure()
        sns.barplot(x="count", y="genre", data=genre_df.sort_values(
            "count", ascending=False))
        plt.title("Distribution of Movie Genres")
        plt.xlabel("Count")
        plt.ylabel("Genre")
        plt.savefig("eda_plots/genre_distribution.png")
        plt.close()

        # Plot budget vs. rating correlation
        plt.figure()
        sns.scatterplot(x="budget", y="avg_rating", data=corr_df)
        plt.title(
            f"Budget vs. Average Rating (Correlation: {correlation:.2f})")
        plt.xlabel("Budget (USD)")
        plt.ylabel("Average Rating")
        plt.savefig("eda_plots/budget_vs_rating_scatter.png")
        plt.close()

        # Plot ratings by year
        plt.figure()
        sns.lineplot(x="release_year", y="rating_count",
                     data=ratings_by_year, label="Rating Count")
        plt.title("Ratings Count by Release Year")
        plt.xlabel("Release Year")
        plt.ylabel("Count")
        plt.savefig("eda_plots/ratings_by_year.png")
        plt.close()

        # Plot average budgets by year
        plt.figure()
        sns.lineplot(x="release_year", y="avg_budget",
                     data=budgets_by_year, label="Average Budget")
        plt.title("Average Budget by Release Year")
        plt.xlabel("Release Year")
        plt.ylabel("Average Budget (USD)")
        plt.savefig("eda_plots/budgets_by_year.png")
        plt.close()

        # Log results
        logging.info("Missing values in ratings_df:\n" +
                     ratings_missing.to_string())
        logging.info("Missing values in merged_df:\n" +
                     movies_missing.to_string())
```

```
        logging.info(
            f"Unique users: {unique_users}, Unique movies (ratings):␣
↪{unique_movies_ratings}, Unique movies (metadata): {unique_movies}")
        logging.info(
            f"Budget vs. average rating correlation: {correlation:.2f}")
        logging.info("Genre distribution:\n" + genre_df.to_string())
        logging.info("Ratings by year:\n" + ratings_by_year.to_string())
        logging.info("Budgets by year:\n" + budgets_by_year.to_string())

        # Save to file
        with open("eda_plots/additional_eda.txt", "w") as f:
            f.write("Missing values in ratings_df:\n" +
                    ratings_missing.to_string() + "\n\n")
            f.write("Missing values in merged_df:\n" +
                    movies_missing.to_string() + "\n\n")
            f.write(
                f"Unique users: {unique_users}\nUnique movies (ratings):␣
↪{unique_movies_ratings}\nUnique movies (metadata): {unique_movies}\n\n")
            f.write(
                f"Budget vs. average rating correlation: {correlation:.2f}\n\n")
            f.write("Genre distribution:\n" + genre_df.to_string() + "\n\n")
            f.write("Ratings by year:\n" +
                    ratings_by_year.to_string() + "\n\n")
            f.write("Budgets by year:\n" + budgets_by_year.to_string() + "\n")

        logging.info("Additional EDA saved to eda_plots/additional_eda.txt")

    except Exception as e:
        logging.error(f"Failed to perform additional EDA: {str(e)}")
        raise


# Call the function
additional_eda(ratings_df, movies_df)
```

```
2025-08-03 19:47:23,363 - INFO - Missing values in ratings_df:
   userId  movieId  rating  timestamp
0       0        0       0          0
2025-08-03 19:47:23,366 - INFO - Missing values in merged_df:
   movieId  title  genres  budget  release_date
0        0      0       0     842           871
2025-08-03 19:47:23,368 - INFO - Unique users: 330975, Unique movies (ratings):
83239, Unique movies (metadata): 50888
2025-08-03 19:47:23,369 - INFO - Budget vs. average rating correlation: 0.09
2025-08-03 19:47:23,376 - INFO - Genre distribution:
                              genre  count
0           Action,Adventure,Fantasy    181
```

1	Comedy,Sport	72
2	Adventure,Family,Fantasy	87
3	Documentary,Western	6
4	Comedy,Drama,Western	6
5	Fantasy,Mystery,Thriller	5
6	Action,Animation,Sci-Fi	41
7	Documentary,Sport	219
8	Fantasy,Sci-Fi,Thriller	4
9	Crime,Documentary,Horror	2
10	Comedy,Family	165
11	Animation,Comedy	38
12	Biography,Documentary,Short	22
13	Action,Comedy	164
14	Animation,Documentary,History	9
15	Drama,Mystery	141
16	Crime	48
17	Biography,Drama,Musical	7
18	Romance	239
19	Comedy,Drama,Mystery	75
20	Adventure,Comedy,Drama	149
21	Drama,Music,Musical	14
22	Action,Drama,Sci-Fi	63
23	Action,Crime	61
24	Biography,Documentary,Sport	85
25	Comedy,Horror,Music	9
26	Thriller	544
27	Adventure,Family,Musical	3
28	Comedy,Drama,History	36
29	Action,Biography,Crime	36
30	Documentary,Short	220
31	Crime,Documentary	135
32	Mystery,Thriller	115
33	Horror,Musical,Short	1
34	Short,Thriller	6
35	Drama,Mystery,Romance	129
36	Sci-Fi,Thriller	58
37	Mystery,Sci-Fi	10
38	Action,Drama,Mystery	45
39	Animation,Comedy,Short	41
40	Crime,Drama,Thriller	570
41	Comedy,Documentary,Music	13
42	Action,Drama,Sport	36
43	Animation,Family,Short	11
44	Adventure,Documentary	38
45	Comedy,Fantasy,Music	5
46	Crime,Mystery	21
47	Crime,Drama,Fantasy	22
48	Animation,Comedy,Drama	103

49	Comedy,Mystery,Thriller	12
50	Documentary,Family,Music	7
51	Adventure,Animation,Fantasy	17
52	Horror,Short	55
53	Comedy,Fantasy,Horror	76
54	Comedy,Documentary,Drama	40
55	Comedy,Mystery,Romance	10
56	Comedy,Sci-Fi,Short	9
57	Adventure,Drama,Romance	43
58	Adventure,Drama,Sci-Fi	38
59	Comedy,Drama,Musical	57
60	Comedy,Drama,Short	71
61	Romance,Sci-Fi,Short	3
62	Comedy,Documentary,Sport	2
63	Adventure,Sci-Fi	10
64	Action,Drama	175
65	Crime,Drama,Romance	124
66	Biography,Comedy,Documentary	56
67	Short,War	4
68	Adventure,Comedy,Fantasy	43
69	Drama,Romance,Sci-Fi	44
70	Adventure,Comedy,Family	152
71	Adventure	33
72	Comedy,Thriller	43
73	Romance,Sport	1
74	Action,Adventure,Short	2
75	Comedy,Crime,Romance	37
76	Comedy,Family,Fantasy	110
77	Drama,Short	233
78	Comedy,Crime,Musical	6
79	Comedy,Documentary,News	3
80	Action,Mystery,Thriller	29
81	\N	43
82	Animation,Fantasy,Short	13
83	Crime,Mystery,Sci-Fi	4
84	Action,Sci-Fi	80
85	Drama,Mystery,Sci-Fi	83
86	Documentary,News	61
87	Comedy,Horror,Thriller	72
88	Action,Talk-Show	1
89	Biography,Short	2
90	Documentary,Musical	9
91	Animation,Drama,Family	39
92	Action,Comedy,Documentary	11
93	Adventure,Comedy,Horror	21
94	Documentary,Horror,Sci-Fi	1
95	Action,Documentary,Drama	5
96	Action,Thriller	273

97	Biography,Documentary,Family	27
98	Documentary,Family,Sci-Fi	1
99	Drama	5107
100	Comedy,Crime,Thriller	49
101	Animation,Short,Western	1
102	Documentary,Music,Musical	5
103	Documentary,Music	462
104	Action,Adventure,Drama	331
105	Animation,Biography,Documentary	17
106	Documentary,Horror	40
107	Documentary,History	284
108	Comedy,Horror,Musical	11
109	Comedy,Fantasy,Short	7
110	War	11
111	Documentary	4003
112	Drama,Musical,Romance	56
113	Fantasy,Short	13
114	Comedy,Crime,Music	5
115	Adventure,Documentary,Western	2
116	Adventure,Drama,History	28
117	Horror,Mystery,Thriller	472
118	Drama,Thriller	684
119	Drama,Romance,Western	9
120	Action,Adventure,Horror	91
121	Action,Drama,Horror	49
122	Drama,History,Thriller	39
123	Documentary,Drama,Sport	11
124	Biography,Drama	304
125	Biography,Crime,Drama	174
126	Comedy,Crime,Mystery	59
127	Drama,Short,War	4
128	Drama,Family,Music	16
129	Action,Crime,History	3
130	Comedy,Romance	1219
131	Adventure,Crime,Drama	27
132	Action,Adventure	31
133	Biography,Drama,Sport	93
134	Drama,Horror,Thriller	271
135	Drama,Sci-Fi,Thriller	96
136	Comedy,Romance,Short	9
137	Drama,Romance	1530
138	Comedy,Mystery	25
139	Biography,Documentary,History	326
140	Crime,Horror	14
141	Comedy,Drama,Thriller	51
142	Action,Comedy,Fantasy	50
143	Comedy,Horror,Romance	22
144	Comedy,Documentary,Fantasy	1

145	Family	132
146	Documentary,Mystery	23
147	Drama,Horror,Romance	29
148	Fantasy	24
149	Crime,Documentary,Mystery	9
150	Comedy,Music,Musical	8
151	Comedy,Family,Sport	18
152	Comedy,Documentary,Family	7
153	Comedy,Drama,Sport	63
154	Crime,Drama	446
155	Biography,Drama,Thriller	18
156	Animation,Family,Fantasy	52
157	Drama,Sport	133
158	Drama,Thriller,Western	16
159	Crime,Sci-Fi,Thriller	6
160	Action,Adventure,Sci-Fi	167
161	Documentary,Drama,Short	11
162	Action,Comedy,War	3
163	Adventure,Animation,Comedy	548
164	Action,Drama,Romance	84
165	Romance,Short	12
166	Family,Fantasy,Romance	12
167	Comedy,Fantasy,Romance	82
168	Crime,Horror,Mystery	66
169	Adventure,Horror,Thriller	14
170	Fantasy,Music,Romance	2
171	Documentary,History,Sport	19
172	Horror,Short,Thriller	12
173	Action,Crime,Horror	22
174	Adventure,Drama	81
175	Documentary,Drama,War	6
176	Comedy,Horror,Short	20
177	Comedy,Fantasy,Sci-Fi	12
178	Biography,Drama,Mystery	11
179	Adventure,Comedy	69
180	Crime,Drama,Musical	9
181	Comedy,Drama,Horror	72
182	Fantasy,Horror,Short	6
183	Comedy,Music,Short	3
184	Adventure,Comedy,Crime	44
185	Action,Drama,History	118
186	Comedy,Drama	2118
187	Adventure,Animation,Documentary	4
188	Biography,Drama,Romance	99
189	Adventure,Comedy,Thriller	4
190	Adventure,Animation,Drama	69
191	Action,Drama,War	74
192	Adventure,Drama,Horror	29

193	Animation,Family	57
194	Crime,Drama,Music	11
195	Drama,Family,Sport	27
196	Adventure,Drama,War	8
197	Comedy,Music,Romance	31
198	Family,Music	5
199	Animation,Horror,Mystery	8
200	Drama,Music,Romance	112
201	Documentary,History,Romance	4
202	Biography,Romance	1
203	Documentary,Music,Short	5
204	Drama,Thriller,War	29
205	Documentary,War	64
206	Biography,Drama,War	26
207	Action,War	16
208	Adventure,Drama,Mystery	24
209	Action,Drama,Thriller	227
210	Horror,Thriller	745
211	Documentary,History,War	60
212	Comedy,History	14
213	Adventure,Animation,Crime	4
214	Adventure,Drama,Family	82
215	Biography,Comedy,Drama	123
216	Drama,Family,Fantasy	50
217	History	16
218	Action,Adventure,Documentary	14
219	Mystery	31
220	Biography,Documentary,Drama	96
221	Drama,Music	134
222	Fantasy,Horror	46
223	Biography,Documentary,Music	184
224	Comedy,Family,Music	18
225	Crime,Drama,History	44
226	Drama,Fantasy,Mystery	79
227	Documentary,Drama,History	45
228	Action,Animation,Drama	56
229	Animation,Crime,Drama	7
230	Action,Horror	46
231	Crime,Documentary,War	2
232	Fantasy,Mystery,Sci-Fi	9
233	Crime,Thriller	157
234	Crime,Drama,Sport	10
235	Action,Crime,Drama	654
236	Action,Drama,Family	11
237	Animation,Drama,Short	33
238	Documentary,History,Music	43
239	Musical	37
240	Comedy,Crime,Horror	24

241	Animation,Horror,Sci-Fi	4
242	Drama,Romance,Thriller	118
243	Crime,Drama,Mystery	574
244	Biography,Comedy	11
245	Animation,Drama,Romance	18
246	Biography,Horror,Thriller	1
247	Action,History,Thriller	5
248	Drama,Fantasy,Thriller	14
249	Animation	59
250	Horror,Thriller,War	4
251	Drama,Horror	128
252	Drama,Fantasy,Horror	182
253	Comedy,Documentary,Horror	5
254	Action,Animation,Fantasy	38
255	Documentary,Fantasy,History	2
256	Drama,Fantasy,Sport	1
257	Horror,Sci-Fi,Thriller	130
258	Adventure,Drama,Thriller	31
259	Action,Adventure,Thriller	55
260	Comedy,Music,Mystery	1
261	Action,Comedy,Crime	310
262	Comedy,Documentary	305
263	Action,Comedy,Horror	125
264	Animation,Biography	3
265	Adventure,Family,History	2
266	Sci-Fi,Short,Thriller	9
267	Adventure,Animation	7
268	Comedy,Drama,Music	210
269	Music	62
270	Action,Comedy,Music	3
271	Action,Adventure,Family	40
272	Drama,Fantasy,Sci-Fi	31
273	Animation,Comedy,Crime	16
274	Comedy,Fantasy	87
275	Animation,Comedy,Family	174
276	Documentary,Drama	136
277	Action,Fantasy	23
278	Music,Romance	4
279	Adventure,Crime,Family	6
280	Drama,Sci-Fi,Short	34
281	Adventure,Family,Sci-Fi	11
282	Action,Horror,Mystery	27
283	Biography,Documentary,News	14
284	Action,Comedy,Sci-Fi	35
285	Adventure,Documentary,Family	10
286	Drama,History,Romance	83
287	Biography,Comedy,Crime	20
288	Family,Fantasy,Horror	6

289	Action,Biography,Drama	60
290	Biography,History	5
291	Adventure,Thriller	5
292	Drama,History,War	110
293	Action,Horror,Thriller	79
294	Horror,Sci-Fi	88
295	Animation,Documentary,Music	1
296	Sci-Fi,Short	77
297	Family,Fantasy	40
298	Drama,Family,Mystery	8
299	Drama,Musical	31
300	Action,Comedy,Romance	36
301	Drama,Romance,War	51
302	Fantasy,Horror,Thriller	44
303	Action,Comedy,Musical	6
304	Adventure,Fantasy,Horror	17
305	Drama,Fantasy,Romance	86
306	Crime,Mystery,Romance	9
307	Biography,Sport	3
308	Drama,Horror,Mystery	401
309	Drama,Family	191
310	Drama,Fantasy	86
311	Biography,Crime,Documentary	71
312	Adventure,Documentary,Drama	24
313	Action,Adventure,Comedy	402
314	Comedy,Musical	35
315	Action,Short	9
316	Action,Thriller,War	11
317	Biography,Documentary,War	5
318	Adventure,Biography,Comedy	9
319	Adventure,Drama,Western	8
320	Biography,Drama,History	319
321	Action,History,War	8
322	Comedy,Romance,Sport	24
323	Family,Romance	11
324	Documentary,History,News	36
325	Action,Comedy,Drama	146
326	Action,Crime,Thriller	327
327	Animation,Fantasy,Horror	12
328	Adventure,Animation,Sci-Fi	5
329	Music,Musical	8
330	Adventure,Documentary,History	8
331	Crime,Thriller,War	1
332	Fantasy,Sci-Fi,Short	8
333	Documentary,Drama,Family	30
334	Biography,Family,Sport	2
335	Drama,Mystery,Thriller	377
336	Adventure,Biography,Drama	44

337	Crime,Documentary,Music	4
338	Drama,History	166
339	Comedy,Musical,Romance	30
340	Drama,Romance,Short	31
341	Animation,Sci-Fi	6
342	Comedy,Drama,Romance	1448
343	Comedy,Music	78
344	Fantasy,Romance	12
345	Horror	930
346	Short	153
347	Horror,Mystery,Sci-Fi	69
348	Crime,Documentary,History	39
349	Family,Mystery,Romance	1
350	Drama,Sci-Fi	111
351	Comedy,Horror	311
352	Action,Fantasy,Horror	43
353	Adventure,Family	45
354	Comedy,Family,Romance	83
355	Animation,Drama,Sci-Fi	12
356	Action,Comedy,Thriller	40
357	Comedy,Crime,Drama	296
358	Documentary,Sci-Fi	22
359	Drama,Family,Romance	80
360	Animation,Drama,Mystery	10
361	Crime,Documentary,Drama	23
362	Documentary,Family	45
363	Action,Drama,Fantasy	65
364	Action,Adventure,History	13
365	Action,Adventure,Animation	443
366	Fantasy,Horror,Mystery	53
367	Comedy,Drama,Fantasy	199
368	Crime,Drama,Horror	123
369	Western	7
370	Biography	34
371	Horror,Mystery,Short	5
372	Crime,Horror,Thriller	66
373	Action,Crime,Mystery	54
374	Drama,Family,Musical	6
375	Comedy,Documentary,History	14
376	Action,Fantasy,Romance	4
377	Comedy,Horror,Sci-Fi	55
378	Adventure,Comedy,Sci-Fi	10
379	Drama,Horror,Sci-Fi	63
380	Biography,Documentary	438
381	Comedy,Crime	192
382	Action,Crime,Documentary	5
383	Action,Animation,Crime	29
384	Action,Biography,Sport	5

385	Action,Sci-Fi,Thriller	110
386	Comedy	3012
387	Biography,Comedy,History	2
388	Animation,Sport	2
389	Adventure,Drama,Fantasy	51
390	Animation,Documentary,News	1
391	Drama,Music,Thriller	5
392	Comedy,Drama,Family	346
393	Action,Mystery,Sci-Fi	17
394	Comedy,Sci-Fi	73
395	Adventure,Documentary,Sport	15
396	Action	179
397	Action,Documentary,Sport	6
398	Sport	11
399	Adventure,Horror	9
400	Comedy,Horror,Mystery	53
401	Action,Biography,History	2
402	Animation,Short	114
403	Animation,Family,Music	2
404	Action,Adventure,Crime	107
405	Adventure,Animation,Family	143
406	Comedy,Drama,Sci-Fi	37
407	Comedy,History,Romance	2
408	Romance,Thriller	21
409	Comedy,Romance,Sci-Fi	20
410	Horror,Sci-Fi,Short	16
411	Mystery,Sci-Fi,Thriller	36
412	Action,Horror,Sci-Fi	97
413	Adventure,Documentary,News	2
414	Drama,War	188
415	Animation,Comedy,Romance	14
416	Adventure,Fantasy,Mystery	9
417	Action,Fantasy,Thriller	9
418	Animation,Family,Musical	13
419	Sci-Fi	93
420	Drama,Musical,War	2
421	Drama,Western	67
422	Action,Drama,Western	21
423	Horror,Mystery	145
424	Biography,Drama,Music	125
425	Family,Musical,Romance	3
426	Comedy,Fantasy,Mystery	4
427	Comedy,Short	130
428	Action,Crime,Romance	7
429	Action,Adventure,Mystery	23
430	Action,Animation,Horror	8
431	Action,Animation	5
432	Animation,Comedy,Fantasy	29

433	Crime,Mystery,Thriller	156
434	Fantasy,Horror,War	1
435	Comedy,Musical,Sci-Fi	3
436	Drama,Romance,Sport	38
437	Documentary,Music,Sport	3
438	Action,Documentary	4
439	Animation,Fantasy,Sci-Fi	7
440	Comedy,Musical,Western	2
441	History,Western	1
442	Animation,Romance,Short	4
443	Animation,Horror,Thriller	2
444	Fantasy,Horror,Sci-Fi	20
445	Adventure,Fantasy,Sci-Fi	7
446	Comedy,Sci-Fi,Thriller	6
447	Documentary,Drama,Romance	7
448	Documentary,Drama,News	11
449	Adventure,Horror,Sci-Fi	11
450	Action,Drama,Short	5
451	Mystery,Romance,Thriller	15
452	Fantasy,Romance,Thriller	3
453	Fantasy,Thriller	11
454	Animation,Drama,Music	6
455	Mystery,Short	2
456	Animation,Biography,Drama	11
457	Animation,Horror,Short	15
458	Fantasy,Horror,Romance	8
459	Animation,Drama,Thriller	2
460	Biography,Documentary,Horror	1
461	Family,Fantasy,Music	2
462	Action,Comedy,Western	1
463	Action,Comedy,Family	17
464	Documentary,Drama,Sci-Fi	4
465	Action,Animation,Comedy	59
466	Adventure,Sci-Fi,Thriller	8
467	Reality-TV	7
468	Drama,Short,Thriller	14
469	Adventure,Comedy,War	2
470	Comedy,Fantasy,Musical	9
471	Adventure,Comedy,History	3
472	Drama,Horror,Short	10
473	Biography,Drama,Family	37
474	Comedy,Crime,Short	3
475	Comedy,Drama,War	23
476	Drama,Music,Short	6
477	Drama,Musical,Sport	1
478	Biography,Documentary,Romance	8
479	Documentary,Family,History	11
480	Animation,Biography,Fantasy	1

481	Comedy,Family,Short	1
482	Action,Comedy,Short	9
483	Adventure,Romance	5
484	Documentary,History,Short	7
485	Comedy,War	7
486	Horror,Romance,Thriller	8
487	Action,Crime,Sci-Fi	11
488	Adventure,Horror,Mystery	13
489	Animation,Drama,Fantasy	49
490	Drama,Horror,Western	7
491	Comedy,Crime,History	2
492	Adventure,Comedy,Western	2
493	Crime,Drama,Western	6
494	Comedy,Crime,Sci-Fi	2
495	Animation,History,Short	1
496	Family,Sport	11
497	Action,Sci-Fi,Short	21
498	Drama,Family,Sci-Fi	8
499	Mystery,Romance,Sci-Fi	2
500	Documentary,Romance	17
501	Drama,Mystery,Short	13
502	Adventure,Comedy,Sport	2
503	Romance,Sci-Fi	6
504	Animation,Mystery,Sci-Fi	5
505	Horror,Mystery,Romance	6
506	Documentary,Drama,Horror	1
507	Action,Adventure,War	3
508	Action,Fantasy,Sci-Fi	14
509	Crime,Documentary,News	3
510	Biography,Music	3
511	Comedy,Documentary,Musical	3
512	Drama,Fantasy,Music	6
513	Drama,Family,History	11
514	Biography,Documentary,Thriller	3
515	Action,Comedy,History	5
516	Family,Fantasy,Musical	5
517	Animation,Comedy,Horror	10
518	Drama,History,Music	11
519	Action,Sport	15
520	Action,Biography,Documentary	10
521	Animation,Drama,History	5
522	Adventure,Biography,Documentary	29
523	Fantasy,Sci-Fi	11
524	Crime,Drama,Short	4
525	Adventure,Mystery,Sci-Fi	5
526	Crime,History,Thriller	6
527	Animation,Fantasy	13
528	Adventure,Comedy,Romance	23

529	Crime,Documentary,Sport	4
530	Drama,Family,Short	11
531	Comedy,Family,Sci-Fi	7
532	Animation,Family,Sci-Fi	4
533	Animation,Horror,Music	1
534	Animation,Documentary	16
535	Drama,Fantasy,Musical	3
536	Comedy,Documentary,Short	2
537	Adventure,Mystery	3
538	Comedy,Family,Musical	18
539	Mystery,Romance	5
540	History,Short	3
541	Animation,Family,History	3
542	Adventure,Fantasy	28
543	Documentary,History,Sci-Fi	3
544	Comedy,Crime,Family	12
545	Drama,Music,Mystery	4
546	Family,Fantasy,Mystery	2
547	Animation,Family,Romance	4
548	Action,Musical,Romance	2
549	Animation,Musical,Short	2
550	Crime,Fantasy,Horror	7
551	Animation,Biography,Comedy	3
552	Animation,Sci-Fi,Short	15
553	Crime,Romance,Thriller	9
554	Action,Sport,Talk-Show	1
555	Animation,Drama,War	1
556	Animation,Sci-Fi,War	1
557	Musical,Short	4
558	Documentary,News,Short	2
559	Adventure,Family,Western	3
560	Drama,History,Sport	8
561	Biography,Documentary,Musical	1
562	Crime,Documentary,Thriller	6
563	Action,Fantasy,Musical	2
564	Adventure,Fantasy,History	2
565	Horror,Romance	6
566	Drama,History,Mystery	12
567	Biography,Comedy,Music	2
568	Horror,War	2
569	Documentary,Short,Sport	7
570	Drama,Mystery,Western	2
571	Drama,Sport,Thriller	5
572	Documentary,Thriller	5
573	Drama,Fantasy,History	7
574	Animation,Comedy,Musical	7
575	Musical,Romance,Short	2
576	Animation,Drama,Sport	3

577	Animation,Comedy,Sci-Fi	12
578	Biography,Drama,Western	1
579	Animation,Fantasy,Mystery	5
580	History,Sport,Thriller	1
581	Family,Horror	1
582	Adventure,Comedy,Music	3
583	Drama,History,Western	6
584	Action,Adventure,Biography	15
585	Action,Horror,Short	2
586	Drama,Horror,Music	6
587	Drama,History,Musical	4
588	Animation,Horror	5
589	Comedy,Crime,Documentary	2
590	Documentary,Sport,Talk-Show	1
591	Animation,Drama,Musical	2
592	Music,Short	4
593	Action,Romance	10
594	Documentary,Drama,Mystery	4
595	Drama,Family,War	3
596	History,War	8
597	Action,Drama,Musical	8
598	Action,Adventure,Romance	6
599	Action,Animation,Family	5
600	Animation,Comedy,Music	6
601	Horror,Romance,Sci-Fi	4
602	Documentary,Fantasy,Mystery	1
603	Documentary,Music,War	4
604	Drama,History,Short	2
605	Adventure,Drama,Sport	3
606	Biography,Thriller	1
607	Action,Romance,Thriller	6
608	Drama,Family,Thriller	6
609	Comedy,Western	6
610	Biography,Drama,Fantasy	5
611	Documentary,Family,Mystery	2
612	Animation,History,News	1
613	Adventure,Music,Sci-Fi	2
614	Animation,Drama	17
615	Animation,Drama,Horror	10
616	Comedy,Fantasy,Thriller	3
617	Comedy,Reality-TV	1
618	Adventure,Crime,Mystery	3
619	Adventure,Mystery,Thriller	3
620	Action,Crime,Musical	2
621	Talk-Show	1
622	Animation,Documentary,Mystery	1
623	Family,Sci-Fi	7
624	Documentary,Mystery,Sci-Fi	1

625	Documentary,Horror,Thriller	3
626	Drama,History,Horror	5
627	Action,Comedy,Mystery	7
628	History,Thriller	1
629	Animation,Comedy,Mystery	4
630	Comedy,Crime,Fantasy	7
631	Biography,Drama,Short	1
632	Action,Romance,Sci-Fi	4
633	Drama,Mystery,War	4
634	Biography,Documentary,Sci-Fi	3
635	Animation,Fantasy,Music	4
636	Adventure,Biography,Crime	3
637	Drama,Music,War	3
638	Short,Sport	2
639	Adventure,Short	2
640	Adventure,Fantasy,Thriller	1
641	Drama,Fantasy,War	2
642	Comedy,Mystery,Sci-Fi	7
643	Family,Music,Musical	5
644	Documentary,Family,Musical	2
645	Animation,Documentary,Drama	3
646	Biography,Crime,Thriller	1
647	Documentary,History,Western	1
648	Adventure,History	6
649	Fantasy,Mystery,Romance	5
650	Fantasy,History,Short	1
651	Animation,Comedy,Talk-Show	1
652	Adventure,History,War	1
653	Comedy,Horror,Western	2
654	Documentary,Music,News	2
655	Action,Animation,Music	2
656	Adventure,Comedy,Documentary	7
657	Adventure,Family,Mystery	11
658	Action,Documentary,History	3
659	Musical,Thriller	1
660	Action,Drama,Music	3
661	Documentary,Fantasy	7
662	Adventure,Animation,Horror	2
663	Animation,Short,War	1
664	Documentary,Short,War	4
665	Documentary,Drama,Thriller	3
666	Comedy,Musical,Mystery	1
667	Animation,Short,Thriller	1
668	Drama,Family,Western	5
669	Biography,Documentary,Fantasy	4
670	Family,Mystery	5
671	Adventure,Drama,Music	3
672	Family,Thriller	1

673	Crime,History	2
674	Animation,Romance	1
675	Action,Biography,Thriller	1
676	Comedy,Documentary,War	3
677	Horror,Music,Thriller	3
678	Adventure,Biography,War	1
679	Adventure,Crime,Horror	2
680	Adventure,Comedy,Musical	3
681	Documentary,Reality-TV	2
682	Crime,Drama,Sci-Fi	8
683	Action,Romance,War	2
684	Documentary,Family,War	1
685	Documentary,Music,Mystery	1
686	Fantasy,Musical,Romance	2
687	Crime,Drama,Family	8
688	Adult,Animation	1
689	History,Mystery,Thriller	1
690	Comedy,Romance,Thriller	8
691	Action,Family	5
692	Comedy,History,Mystery	1
693	Romance,Short,War	1
694	Fantasy,Short,Thriller	1
695	Action,Animation,Sport	1
696	Drama,Musical,Thriller	3
697	Game-Show	1
698	Family,Short	5
699	Action,History	3
700	Comedy,Talk-Show	2
701	Animation,History	1
702	Animation,History,Music	1
703	Action,Crime,Fantasy	10
704	Comedy,Documentary,Thriller	1
705	Adventure,Animation,Biography	2
706	Documentary,Horror,Mystery	3
707	Music,Sci-Fi	2
708	Adventure,Family,Romance	1
709	Adventure,Comedy,Mystery	2
710	Drama,Sci-Fi,Sport	1
711	Action,Comedy,Reality-TV	2
712	Family,History,War	1
713	Adventure,History,Mystery	1
714	Adventure,Animation,Short	3
715	Horror,Musical	3
716	Documentary,Family,News	1
717	History,Sport	2
718	Drama,Music,Sci-Fi	2
719	Documentary,Drama,Fantasy	5
720	Animation,History,Horror	1

721	Comedy,Family,Mystery	5
722	Horror,Romance,Short	1
723	Documentary,Drama,Music	8
724	Biography,Fantasy,Short	1
725	Animation,Family,Mystery	1
726	Biography,Comedy,Sci-Fi	1
727	Drama,Fantasy,Short	9
728	Drama,Musical,Short	2
729	Animation,Comedy,Documentary	4
730	Fantasy,Music,Sci-Fi	1
731	Fantasy,Romance,Sci-Fi	1
732	Action,Sport,Thriller	1
733	Biography,Drama,Sci-Fi	1
734	Documentary,Fantasy,Music	1
735	Drama,Family,Horror	1
736	Drama,Horror,Musical	5
737	Action,Family,Horror	1
738	Adventure,Fantasy,Romance	6
739	Crime,Short,Thriller	2
740	History,Mystery,Sci-Fi	1
741	Musical,Romance	3
742	History,Sci-Fi,Thriller	1
743	Animation,Documentary,Short	4
744	Biography,Drama,Horror	2
745	Animation,Fantasy,Musical	3
746	Comedy,Documentary,Romance	4
747	Comedy,Horror,Sport	1
748	Crime,Romance	1
749	Comedy,Family,Horror	3
750	Action,Music	1
751	Short,Western	2
752	Crime,Horror,Sci-Fi	5
753	Action,Fantasy,Mystery	3
754	Adventure,Crime,Documentary	1
755	Crime,Horror,Romance	3
756	Animation,Documentary,War	1
757	Horror,Musical,Sci-Fi	2
758	Animation,Crime,Documentary	1
759	Fantasy,War	1
760	Adventure,Documentary,Short	2
761	Family,Mystery,Sport	1
762	Animation,Romance,Sci-Fi	2
763	History,Horror,Thriller	2
764	Crime,Documentary,Short	2
765	Adventure,Animation,Music	1
766	Adventure,Documentary,Romance	1
767	Documentary,History,Mystery	1
768	Animation,Mystery	1

769	Adventure,Sci-Fi,Short	4
770	Family,Musical,Sci-Fi	1
771	Family,Romance,Sport	1
772	Comedy,Short,Sport	1
773	Animation,Biography,Short	1
774	Biography,History,War	3
775	Comedy,Fantasy,Reality-TV	1
776	Action,Fantasy,Short	1
777	Action,Documentary,Short	1
778	Fantasy,Mystery,Short	1
779	Action,Sci-Fi,Sport	2
780	Action,Biography,Comedy	5
781	Animation,Family,Sport	1
782	Comedy,Music,War	1
783	Horror,Music,Short	2
784	Adventure,Fantasy,Musical	1
785	Biography,Music,Romance	1
786	Action,Romance,Short	1
787	Action,Music,Thriller	1
788	Comedy,Documentary,Mystery	1
789	Action,Family,Fantasy	2
790	Musical,Sci-Fi	1
791	Drama,Short,Western	3
792	Drama,Sci-Fi,War	1
793	Action,Drama,News	1
794	Comedy,Fantasy,History	2
795	Action,Western	1
796	Action,History,Romance	1
797	Action,Animation,Romance	1
798	Comedy,Crime,Sport	3
799	Biography,History,Romance	1
800	Animation,Sci-Fi,Thriller	1
801	Horror,Thriller,Western	1
802	Romance,Sci-Fi,Thriller	4
803	Comedy,Short,Western	1
804	Crime,Horror,Musical	1
805	Animation,Biography,Crime	1
806	Animation,Mystery,Short	1
807	Action,Comedy,Sport	4
808	Documentary,History,Horror	5
809	Family,Musical	2
810	Comedy,History,Musical	1
811	Biography,Fantasy,Sci-Fi	1
812	Fantasy,Music	1
813	Sport,Thriller	1
814	Documentary,News,War	1
815	Action,Animation,Short	2
816	Comedy,Family,History	2

817	Crime,Short	2
818	Documentary,History,Thriller	3
819	Crime,Fantasy,Mystery	1
820	Comedy,Musical,Short	2
821	Comedy,History,War	2
822	Action,Adventure,Musical	2
823	Sci-Fi,War	1
824	Documentary,Horror,Romance	1
825	Adventure,Fantasy,Short	1
826	Drama,History,Sci-Fi	3
827	History,Romance	1
828	Adventure,Animation,History	1
829	Biography,Fantasy,Horror	1
830	Adventure,Documentary,Mystery	1
831	Crime,Horror,Music	1
832	Animation,Documentary,Sci-Fi	1
833	Family,Fantasy,History	1
834	Adventure,Sport	2
835	Drama,Horror,War	1
836	Crime,Romance,Sci-Fi	1
837	Documentary,Horror,Short	2
838	Documentary,Family,Fantasy	2
839	Adventure,Documentary,Sci-Fi	1
840	Biography,Musical	1
841	History,Romance,War	1
842	Family,Music,Romance	2
843	Adventure,Musical	1
844	Crime,Thriller,Western	1
845	Crime,Family,Mystery	2
846	Fantasy,Music,Musical	1
847	Action,Family,Romance	1
848	Horror,Musical,Mystery	1
849	Reality-TV,Short	1
850	Action,Musical,Thriller	1
851	Animation,Musical	1
852	Crime,Drama,War	3
853	Adult,Fantasy	1
854	Documentary,Family,Sport	1
855	Documentary,Family,Short	2
856	Comedy,Crime,Western	1
857	Adventure,Family,Music	1
858	Thriller,Western	1
859	Action,Horror,War	2
860	History,Horror	1
861	Action,Documentary,Mystery	1
862	Comedy,Fantasy,Sport	1
863	Comedy,Sport,Thriller	1
864	Action,Family,Mystery	1

```
865    Biography,Documentary,Mystery      3
866           Biography,Thriller,War      1
867                Action,Biography      1
868            Family,Fantasy,Short      1
869                 Animation,Music      1
870       Animation,Fantasy,Thriller      1
871           Mystery,Short,Thriller      1
872            Horror,Music,Mystery      3
873           Fantasy,Musical,Sci-Fi      1
874          Animation,Crime,Mystery      2
875              Comedy,Music,Sci-Fi      1
876         Musical,Romance,Thriller      1
877            Animation,Music,Short      1
878        Documentary,Music,Romance      1
879       Documentary,Fantasy,Horror      1
880           Animation,Music,Sci-Fi      1
881              Drama,News,Thriller      1
882             Action,Romance,Sport      1
883             Mystery,Sci-Fi,Short      1
884             Sci-Fi,Short,Western      1
885        Documentary,Drama,Musical      1
886                  Fantasy,Mystery      1
887      Adventure,Animation,Musical      1
888           Crime,Fantasy,Thriller      1
889           Family,Sci-Fi,Thriller      1
890            Action,Horror,Western      1
891              Drama,Mystery,Sport      1
892                     Adult,Comedy      1
893                     Adult,Horror      1
894        Action,Animation,Thriller      1
895                      Music,Sport      1
896               Adult,Comedy,Music      1
897           Fantasy,Horror,Musical      1
898     Animation,Documentary,Horror      1
899         Animation,Horror,Musical      1
900         Action,Animation,Mystery      1
901            Drama,Musical,Mystery      1
902              Crime,Music,Mystery      1
903          Adventure,Crime,History      1
904                Drama,Short,Sport      1
905          Comedy,Musical,Thriller      1
906                    Crime,Fantasy      1
2025-08-03 19:47:23,382 - INFO - Ratings by year:
    release_year  rating_count  avg_rating
0         2025.0             8    2.062500
1         2003.0        976764    3.486380
2         2007.0        783460    3.512716
3         2018.0        279346    3.503675
```

```
4          2015.0        476129    3.578361
5          2023.0         10361    3.281006
6          2006.0        785797    3.546563
7          2022.0         69804    3.365065
8          2013.0        533465    3.498900
9             NaN         63727    3.835141
10         1997.0            23    3.543478
11         2014.0        597280    3.641847
12         1979.0             5    3.400000
13         2019.0        249724    3.562133
14         2004.0       1055793    3.517918
15         1998.0             8    3.250000
16         2020.0         93748    3.328983
17         2012.0        555773    3.553292
18         2009.0        689221    3.533179
19         2016.0        423352    3.548732
20         2001.0       1180764    3.509128
21         2024.0           139    2.964029
22         2005.0        740893    3.446953
23         2000.0       1232002    3.429360
24         2010.0        659697    3.572332
25         2011.0        546446    3.489451
26         2008.0        739024    3.525120
27         2017.0        341912    3.545809
28         1999.0            14    3.428571
29         2002.0       1052123    3.488750
30         2021.0         95868    3.350237
31         1978.0             1    3.000000
32         1994.0            15    3.266667
33         1972.0             1    3.000000
34         1996.0             3    2.833333
35         1989.0             4    1.625000
36         1951.0             1    3.000000
37         1984.0             1    3.000000
38         1977.0             1    2.500000
39         1987.0             2    1.500000
2025-08-03 19:47:23,387 - INFO - Budgets by year:
    release_year    avg_budget
0         2003.0  2.326385e+07
1         2007.0  1.860911e+07
2         2018.0  1.907370e+07
3         2015.0  1.752681e+07
4         2023.0  4.140976e+07
5         2006.0  1.845769e+07
6         2022.0  2.816597e+07
7         2013.0  1.829247e+07
8            NaN  5.027625e+04
9         2014.0  1.690592e+07
```

```
10        2019.0  2.036461e+07
11        2004.0  2.264225e+07
12        2020.0  1.393325e+07
13        2012.0  1.892594e+07
14        2009.0  1.932364e+07
15        2016.0  1.954207e+07
16        2001.0  2.487622e+07
17        2005.0  2.188737e+07
18        2000.0  2.630241e+07
19        2010.0  2.009321e+07
20        2011.0  1.833702e+07
21        2008.0  1.862202e+07
22        2017.0  1.878238e+07
23        2002.0  2.232461e+07
24        2021.0  2.669242e+07
25        2024.0  3.031494e+07
26        2025.0  3.000000e+07
27        1999.0  1.200000e+05
2025-08-03 19:47:23,402 - INFO - Additional EDA saved to
eda_plots/additional_eda.txt
```

[7]:
```python
# Cell 8: Clean Up
def cleanup():
    try:
        spark.stop()
        logging.info("Spark session closed.")
    except Exception as e:
        logging.error(f"Cleanup failed: {str(e)}")
        raise


cleanup()
```

```
2025-08-03 19:48:00,360 - INFO - Spark session closed.
```