IMDB Movie Database

ANALYSIS OF IMDB DATABASE

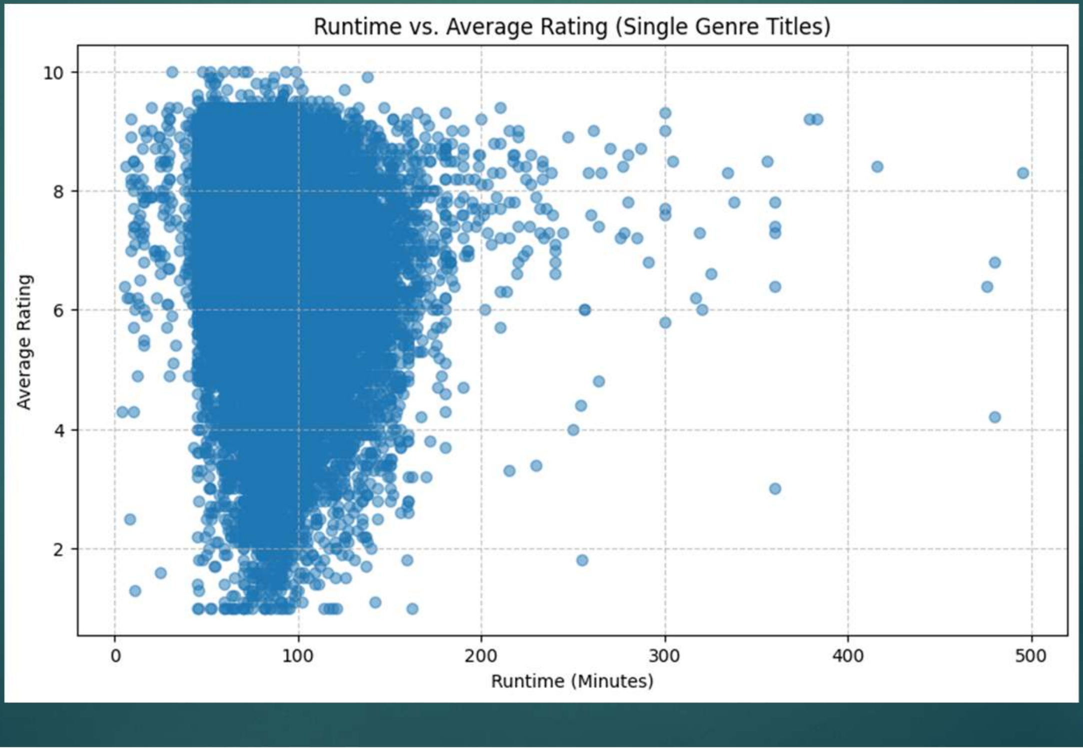
Database Overview

Upon my initial examination of the database, I found that it consists of 8 tables, ranging in all shapes and sizes. A vast array of data is represented within them relating to movies released between 2010 and 2019, the data found lies within two broad categories, principals and movies.

The movie portion of the database consists of over 960,000 titles. There are several different types of principals listed in that portion, some of the unique principals include: actors, directors, producers, cinematographers, composers, writers and editors.

Business Question: Runtime Vs. Rating

There does appear to be a correlation between runtime and average rating among single genre titles. The higher rated titles seem to be clustered in the 90 – 120-minute area as illustrated below:



Potential Data Cleaning Tasks

Upon thorough examination of the tables in the database, recommendations for cleaning are as follows:

- Convert data types (years as integers, ratings are floats).
- Split genres into multiple fields.
- Handle missing values.
- Remove duplicates.
- Address special characters in fields.