### ****Dataset Overview****

The dataset contains **486 entries** with the following columns:

1. **Movie Name**: Name of the movie (2 missing values).
2. **Rating**: IMDb rating of the movie (2 missing values).
3. **Year**: Year of release (2 missing values).
4. **Duration (min)**: Duration of the movie in minutes (2 missing values).
5. **Actors**: Key actors in the movie (2 missing values).
6. **Genre**: Genre(s) of the movie (2 missing values).

### ****Plan for Analysis****

1. **Data Preprocessing**:
   1. Handle missing values.
   2. Identify and remove duplicates.
   3. Ensure correct data types.
2. **Exploratory Data Analysis (EDA)**:
   1. Summary statistics of numerical columns.
   2. Visualize distributions of key features.
   3. Correlation analysis.
   4. Analyze categorical variables like genres and actors.
3. **Visualization**:
   1. Boxplots, scatterplots, and bar charts.
   2. Heatmap for correlations.
4. **Machine Learning Models**:
   1. Train a regression model to predict movie ratings.
   2. Evaluate the model's performance.

Let me begin by preprocessing the data. ​​

### ****Data Preprocessing Steps****

1. **Duplicate Removal**:
   1. Found **16 duplicate rows** and removed them.
2. **Handling Missing Values**:
   1. Filled missing values in numerical columns:
      1. **Rating**, **Year**, and **Duration** were imputed with their respective means.
   2. Rows with missing values in **Movie Name**, **Actors**, and **Genre** were removed.
3. **Final Dataset**:
   1. Total rows: **469**.
   2. No missing values remain.

### ****Exploratory Data Analysis (EDA)****

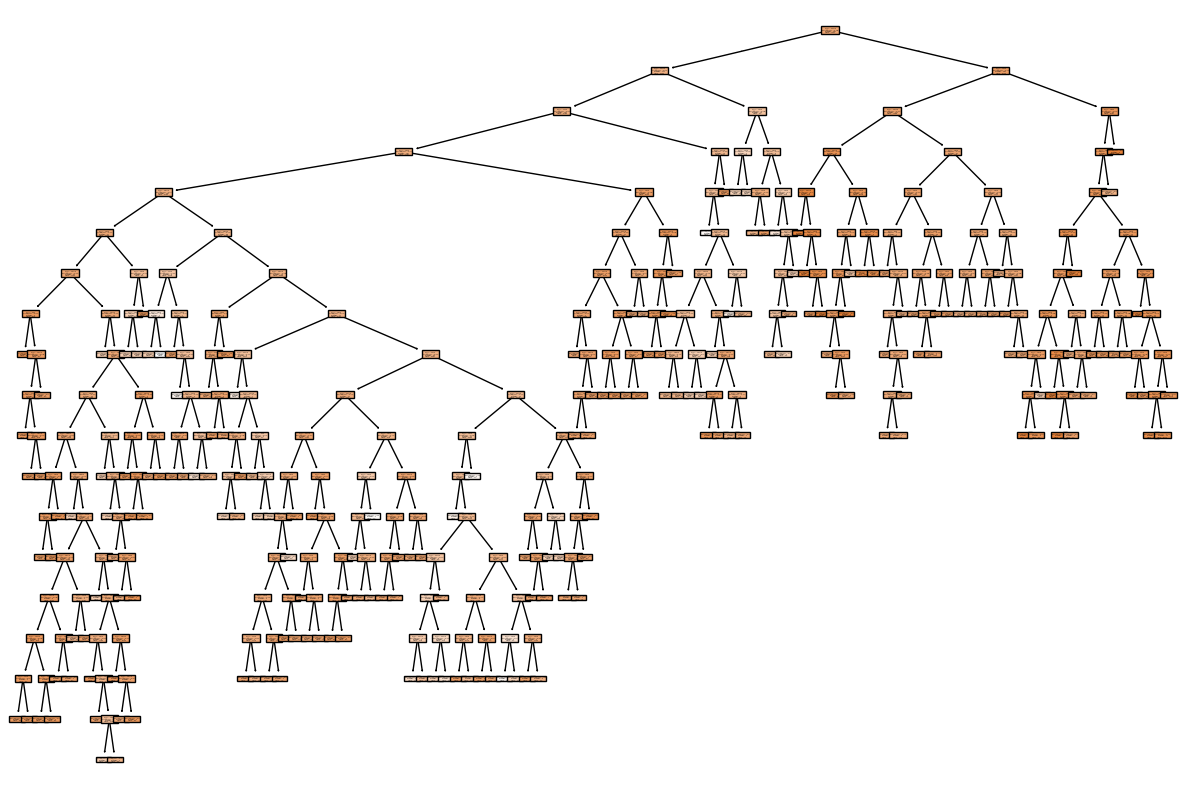
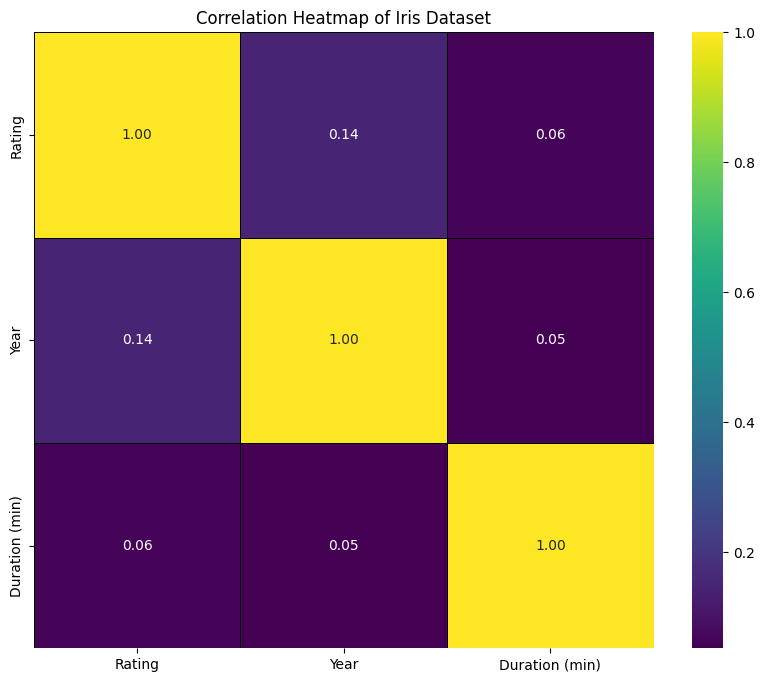
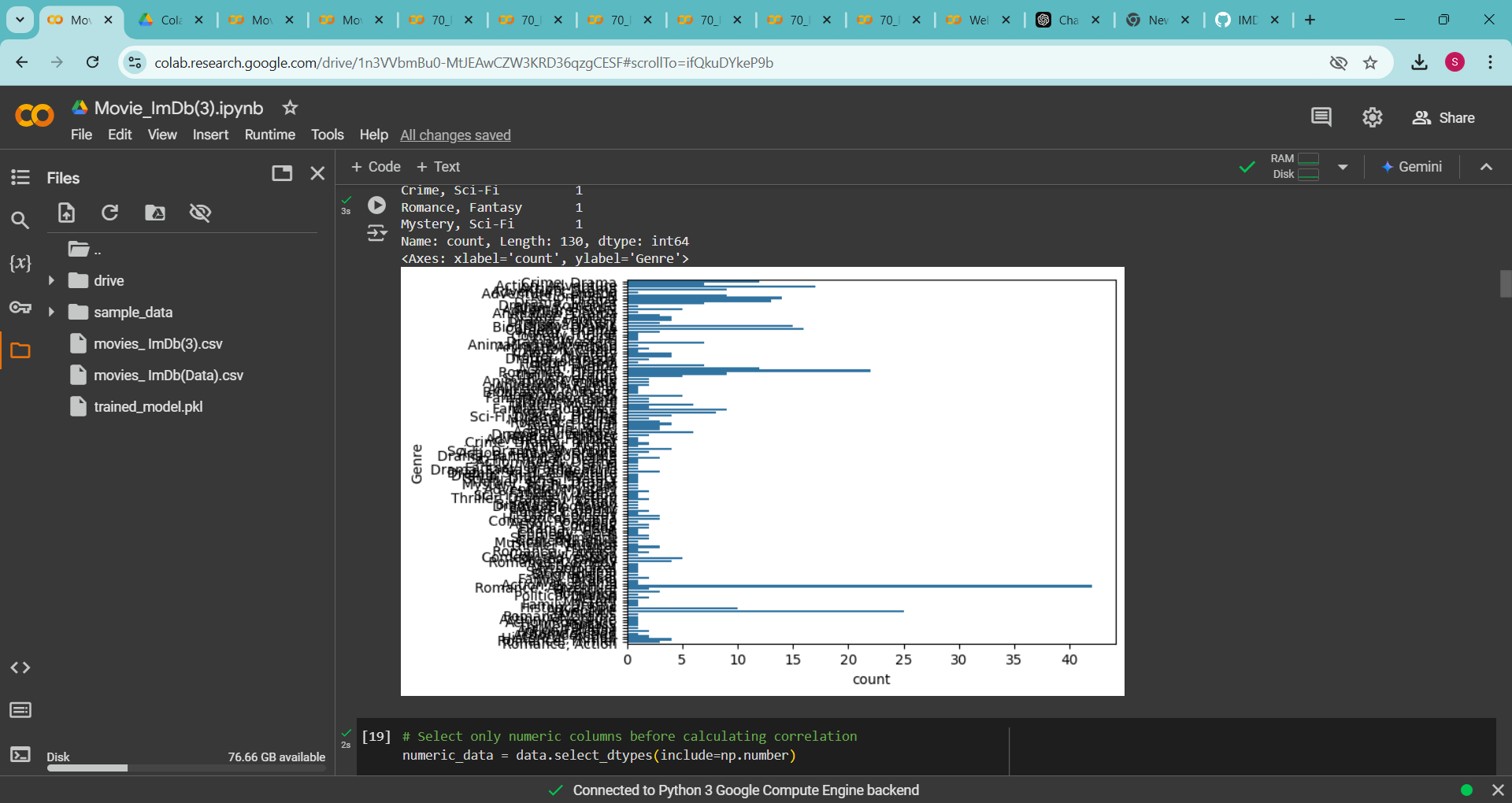
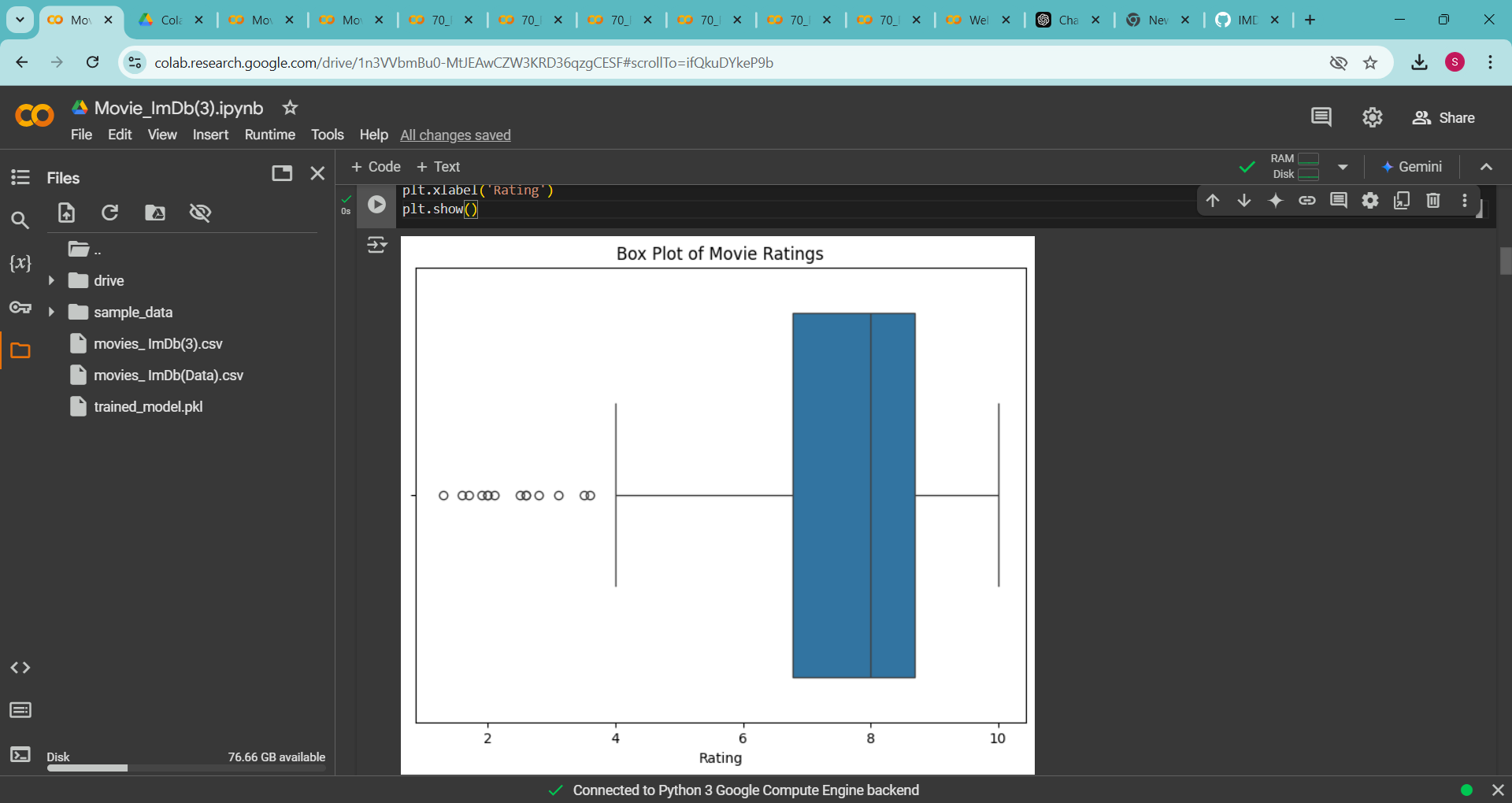
#### ****Summary Statistics for Numerical Columns****

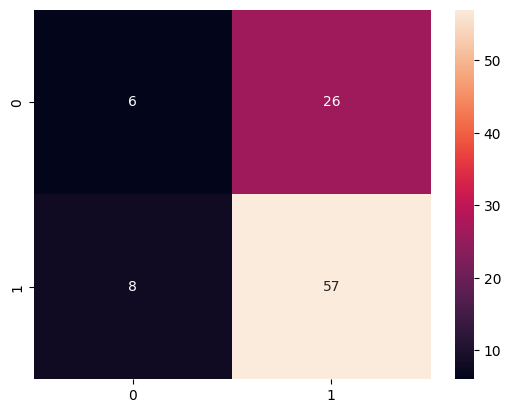
* **Ratings**:
  + Range: 1.3 to 10.
  + Average: 7.49.
  + Most movies are rated between 6.5 and 8.8.
* **Year**:
  + Range: 1940 to 2025.
  + Most movies were released after 2011.
* **Duration**:
  + Range: 81 to 202 minutes.
  + Average: 140 minutes.

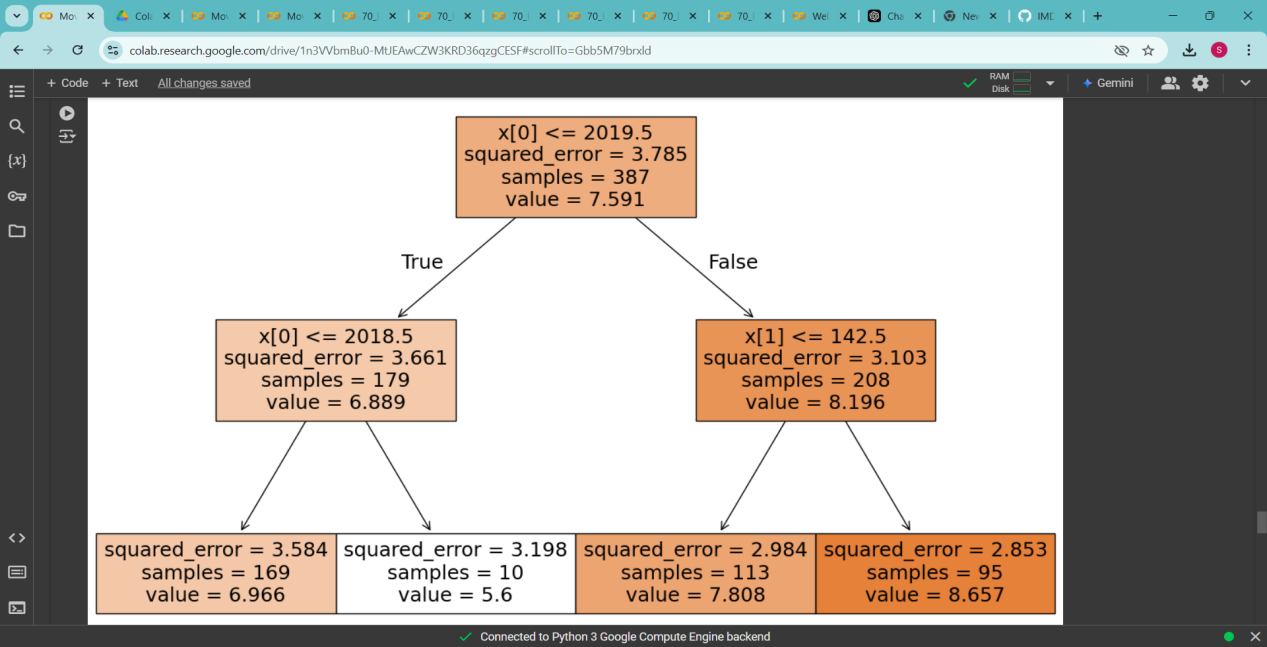
#### ****Observations from Graphs****

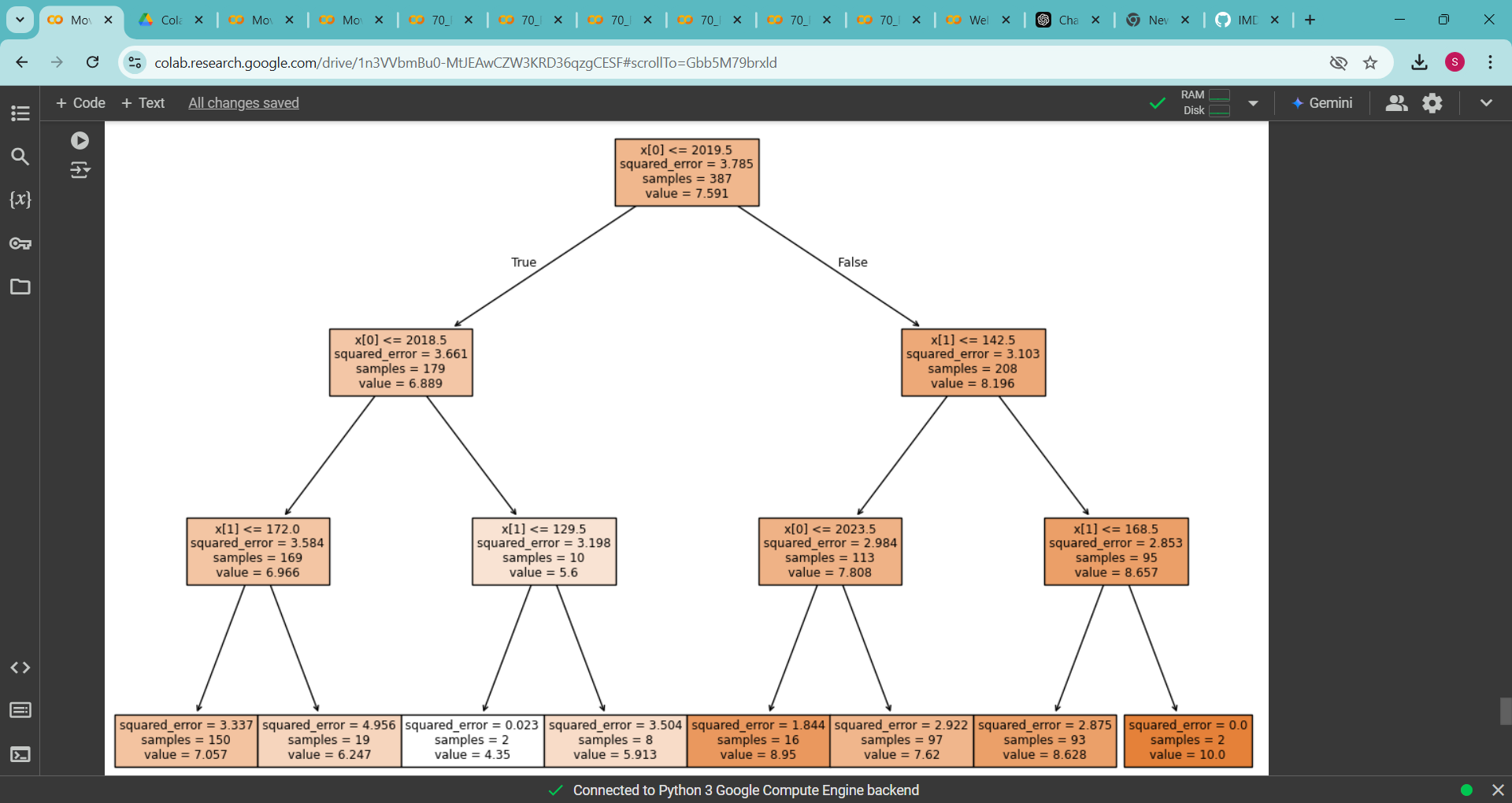
1. **Rating Distribution**:
   * Ratings follow a skewed distribution, with most movies scoring between 6 and 9.
2. **Year Distribution**:
   * A large concentration of movies comes from recent years (2010 onward).
3. **Duration Distribution**:
   * Most movies are between 130 and 150 minutes long.

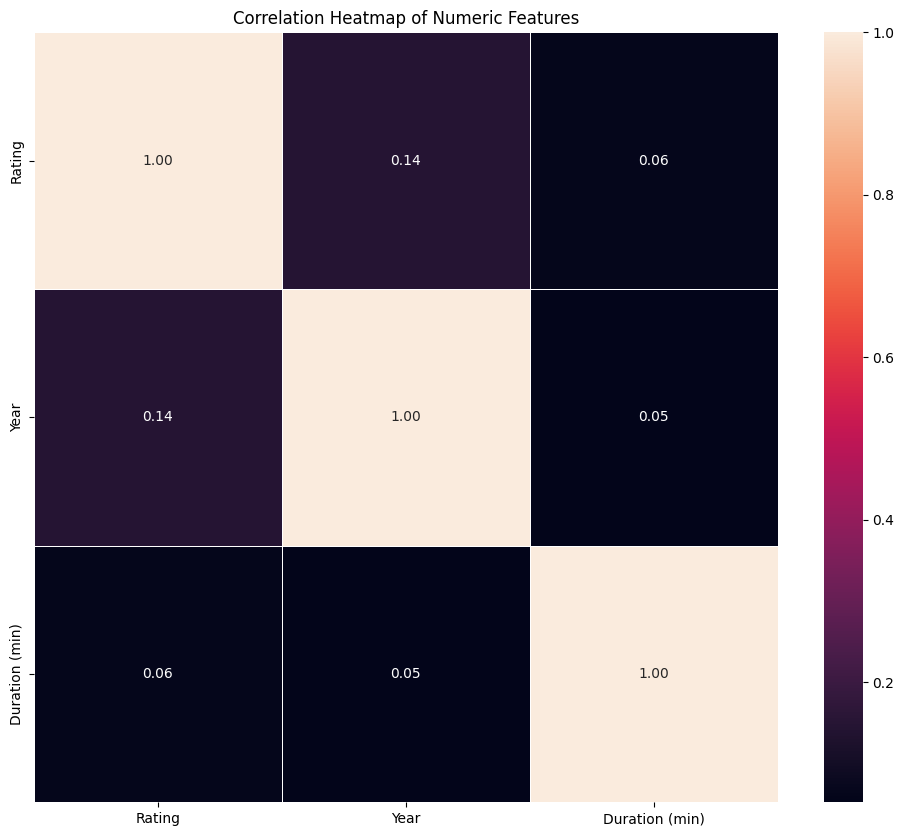
Next, I'll analyze correlations and relationships between features. ​











Output:

