R Project:
Film DataSet

Team 11

Dataset overview

- 31 variables, 3555 observations
- Categorical data

+ Large dataset

- Data diversity

Variables: Closer Look (link)

From 31 Variables, 4 are “unique” (the rest are the film’s genre), namely:

- Country of production: France has the largest share of film produced (rest of films is ONLY co-production between countries)
- Revenue: from hundreds to max 20MM (outlier)
- Year: it spans from 1996 to 2010
- Genre: Drama (1940 movies) and Comedy (1011 movies) have the largest share of the sample
- Aspect Ratio (didn’t look at what it is exactly)

Suggested Approach

Approach:

1. Perform analysis* for entire sample, i.e benchmark (incl. movie genre analysis)
2. Perform analysis for France (based on same code as 1.)
3. Compare benchmark and France
4. Comment on findings with theoretical explanations to sustain our numbers (or state the dataset is nonsense)

What to Include

*The analysis should include:

- Descriptive statistics and plotting for each variable (look for most interesting movies genres)
- Regression with revenues as dependant variable
- T-test, f-test, proportion test (where applicable)

Example of Questions to Answer/ To Do

- Plot revenues over years, revenues per genre, number of genre per year
- Evolution of genres produced over years
- Evolution of revenues over years
- Which genre generate the most revenue
- Why less drama in 2010 than 1996