Правительство Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
«Национальный исследовательский университет
«Высшая школа экономики»

Анализ количества убийств в фильмах

The Analysis of On-screen Movie Kills

Homework Project 2018/2019

Команда:
Борис Цейтлин
Константин Ромащенко
Юлия Гурова

1 курс магистратуры
Программа: «Науки о данных»
Факультет компьютерных наук

Москва 2018

THE RUSSIAN GOVERNMENT
FEDERAL STATE AUTONOMUS EDUCATIONAL INSTITUTION
FOR HIGHER PROFESSIONAL EDUCATION
NATIONAL RESEARCH UNIVERSITY
“HIGHER SCHOOL OF ECONOMICS”

The Analysis of On-screen Movie Kills

Homework Project 2018/2019

The team:
Boris Tseitlin
Konstantin Romashchenko
Yulia Gurova

MSc Program “Data Science”
1^{st} year
Faculty of Computer Science

Moscow 2018

Table Of Contents

1 THE CHOICE OF THE DATASET

The dataset that is used for the project contains information about on-screen deaths in movies. There are 545 movies (more then 100 objects) and 8 characteristics including names, so the dataset meets the requirements.

The sourse for the data is the thematic web-site moviebodycounts.com. This dataset was processed and published on figshare.com. It was gathered in accordance with the rules, which are published on the web-site. We took several characteristics for the consideration: the year of the film, MPAA rating (Motion Picture Association of America film rating system), genre or genres, the name of the director, the lenth of the film in minutes, IMDB rating based on user ratings. The main feature that we consider is the number of on-screen deaths in the movie.

The analysis of the data may reveal how the ratings depend on the number of deaths, how this number is changing with the year of the release and so on. Moreover, genre and length of the film combined with the violence on the screen may give the idea of the age ratings. This is a good set for the classification and clustering problems, as the films are grouped on genres, MPAA ratings which are similar inside the groups, but dissimilar between them.

This study analysis may be the first step to the automation of age rating systems. Moreover it may be helpful in the development of the recommendation systems. And, as watching movies is the common interest of our team, the work with the dataset will inspire us to further conquest in Data Analysis.

2 K-means clustering

3 Cluster Interpretation

4 Contingency Table Analysis

5 PCA: Hidden Factor & Data visualization

6 2D regression