SCHOOL OF INFORMATION AND PHYSICAL SCIENCES

INFT6201 - BIG DATA
LAB PROJECT 2
(PREPARATION)

This lab project covers the lecture material from Weeks 1-5. Please make yourself familiar with the lecture content and the corresponding jupyter notebooks before proceeding. We will be using the CSM movie dataset courtesy of Mehreen Ahmed on the UCI Machine Learning Repository.

PREPARATION

In preparation for lab project 2, load the moviedata.csv dataset into your Jupyter notebook and use it to practice the topics that we discussed in the lecture up to this point. In the Week 4 lab, practice the following concepts:

- Load and explore the *.csv datasets
- Bar charts, box plots, and violin plots
- Adding confidence intervals to box plots
- Using colour palettes to maximize insight
- Determine outliers and extreme values
- Apply trimming and winsorising
- Histograms and density plots
- Scatterplots with two or three variables
- Sankey diagrams
- Parametric and non-parametric tests
- Testing for variance homogeneity

The specification for lab project 2 will be released separately at the start of Week 5.

REFERENCES

Ahmed M, Jahangir M, Afzal H, Majeed A, Siddiqi I (2015). Using Crowd-source based features from social media and Conventional features to predict the movies popularity IEEE International Conference on SmartCity (pp. 273-278).

DATASET

moviedata	*Conventional and Social Media Movies 2014 and 2015*

Description

A dataset about the success of movies in 2014 and 2015.

Usage

moviedata

Format

A data frame with 231 observations on the following 14 variables.

`movie`	Name of the movie
`year`	Year of movie release
`ratings`	Rating of the movie (0 – 10)
`genre`	Identifier for the genre of the movie (e.g., action, adventure, drama)
`gross`	Gross world-wide income from the movie (in US$)
`budget`	Budget for the movie
`screens`	Number of screens that the movie was initially launched in on the opening weekend in the US
`sequel`	A number indicating whether the movie is sequel or original (individual) movie, where higher numbers indicate later sequels in a series. For instance, for Mission Impossible a sequel value of 5 indicates that this is the fifth movie in the series.
`dummy_sequel`	0 – Original movie 1 – Sequel movie
`sentiment`	A sentiment score assessed through an analysis of tweets about the movie on Twitter. 0 represents a neutral sentiment, a positive value represents a positive sentiment, and a negative value indicates a negative sentiment. The sentiment score for each movie was calculated by retrieving all tweets related to each movie, assigning the sentiment score to each of them and then aggregating the score.
`views`	Number of times the movie trailer was viewed on YouTube
`likes`	Number of likes the movie trailer received on YouTube
`dislikes`	Number of dislikes the movie trailer received on YouTube
`comments`	Number of times the movie trailer received a comment on YouTube
`aggregate_followers`	The aggregate number of actor followers: Equal to sum of followers of top 3 cast from Twitter

Source

Ahmed M, Jahangir M, Afzal H, Majeed A, Siddiqi I. Using Crowd-source based features from social media and Conventional features to predict the movies popularity. In Smart City/ SocialCom/S ustainCom (SmartCity), 2015 IEEE International Conference on 2015 Dec 19 (pp. 273-278). IEEE. https://ieeexplore.ieee.org/document/7463737

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.