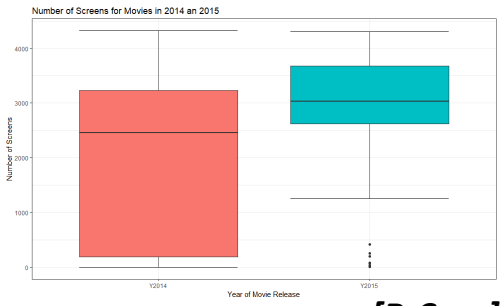
SCHOOL OF ELECTRICAL ENGINEERING AND COMPUTING

BIG DATA & DATA ANALYTICS
LAB PROJECT 2

This lab project is based on a dataset about movie success in 2014 and 2015 by Ahmad et al. (2015) which is available on the online platform by Lichman et al (2013). Download the file movidata.csv from Blackboard and then complete the following exercises.

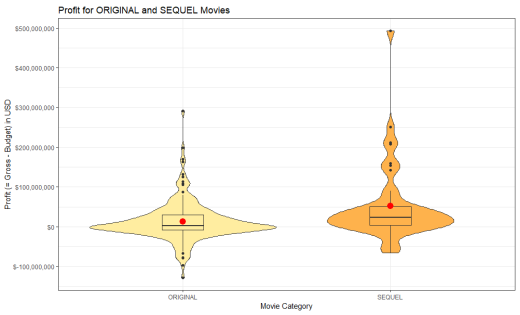
EXERCISE 1 (1 MARK) *[R-CODE]*

Use ggplot() to create a box plot that shows the number of screens on which each movie was initially launched in the US on the y-axis separately for 2014 and 2015. Note: Only include those observations that **do not** have a missing value (NA) for the variable "screens" (e.g., by using !is.na(…)).



EXERCISE 2 (2 MARK) *[R-CODE]*

Calculate the profit of each movie (profit = gross – budget) and add the results as a new variable "profit" to the moviedata dataframe. Use ggplot() to create a violin plot that shows the profit on the y-axis separately for ORIGINAL movies and SEQUEL movies (using the sequelcat variable). Use the "YlOrRd" colour palette from the RColorBrewer library to fill the violin plots (hint for spelling: YlOrRd stands for Yellow / Orange / Red). Add a boxplot on top of the violin plot and add a red point that indicates the mean value. Note: Only include those observations that **do not** have a missing value (NA) for the variable "profit".



EXERCISE 3 (1 MARK) *[R-CODE]*

Use the subset() command to create a subset of the dataframe that only includes observations without missing values for budget, screens, and aggregate_followers. Name this data frame "moviedatasub". Then, using the newly created data frame " moviedatasub ", use the custom winsor() function discussed in the lecture slides in week 3 to create a new variables likes_winsor based on the variable likes. Use a multiplier of 1.5.

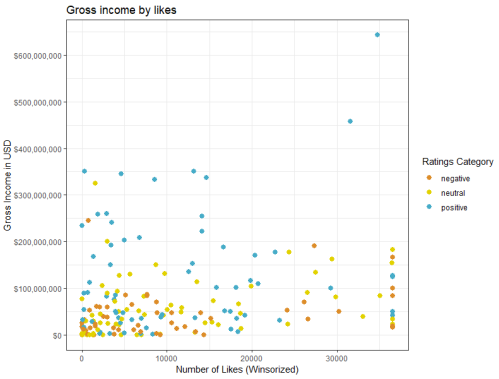
To make sure that the winsorising worked, compare the two variables by creating simple box plots using the following commands.

```
with(moviedatasub, boxplot(likes))
with(moviedatasub, boxplot(likes_winsor))
```

EXERCISE 4 (2 MARKS) *[R-CODE]*

Look up the "cut" command. Based on the dataset "moviedatasub", create a new column "ratingscat" in the dataframe that describes the ratings category of a movie using the cut command. Distinguish between the following categories:

- "negative" (0 ≤ rating < 6)
- "neutral" (6 ≤ rating < 6.8)
- "positive" (6.8 ≤ rating < 10)



Use ggplot() to create a scatterplot for gross_winsor over likes_winsors that you created in Exercise 3. Indicate the different ratings categories by colouring the points in the scatterplot with the "FantasticFox" color palette of the "wesanderson" library package.

EXERCISE 5 (1 MARK) *[R-CODE]*

Based on the dataset "moviedatasub", use the ddply() function of the package "plyr" to create a data frame with the means and standard deviations of profit, gross, and budget for the three different ratings categories (variable: ratingscat, cf. Exercise 4) and for the two different values of sequelcat (ORIGINAL / SEQUEL). Also include the number of observations N for each of the category combinations. The output should look like this:

	ratingscat	sequelcat	N	profit_avg	profit_sd	gross_avg	gross_sd	budget_avg	budget_sd
1	negative	ORIGINAL	41	-272956.8	46907791	39060336	42021960	39333293	42816923
2	negative	SEQUEL	10	22003000.0	19465550	60253000	71073044	38250000	62365077
3	neutral	ORIGINAL	50	8570419.0	40523417	46181075	50940229	37610656	41910731
4	neutral	SEQUEL	18	34738888.9	72196384	98116667	72482608	63377778	34388499
5	positive	ORIGINAL	54	33555944.4	61828502	83507796	88925279	49951852	52966518
6	positive	SEQUEL	14	98314285.7	140955415	266528571	143007167	168214286	50710610

EXERCISE 6 (2 MARKS) *[R-CODE]*

Based on the dataset "moviedatasub", use a Bartlett's test to test for variance homogeneity in the variable profit across the three different ratings categories (variable: ratingscat, cf. Exercise 4). In your own words, interpret the results of the test and decide whether we should assume that the variances are homogeneous.

Then, use a one-way Analysis of Variance (ANOVA) to test whether there is a difference in mean profit across the three different ratings categories and interpret the result in your own words. Conduct a PostHoc analysis to determine which groups are significantly different from each other. How does the result of the test of variance homogeneity affect the PostHoc analysis?

EXERCISE 7 (1 MARKS) *[R-CODE]*

Based on the dataset "moviedatasub", compare the mean profits for ORIGINAL and SEQUEL movies (variable: sequelcat). Which test should we use to test whether there is a significant difference and why? Conduct the test in R and interpret the result in your own words.

REFERENCES

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.

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.