# Статистика и емпирични методи

## Контролно №1

## Група 4

Задачата за контролното е да се изследват данни за 3801 филма. Данните са във файла `movie.csv`, който е с header, и трябва да бъдат свалени от тук. Записаните променливи са следните:

- `color`: Дали филмът е цветен или черно бял;
- `num_critic_for_reviews`: Брой коментари на критици;
- `duration`: Продължителност на филма;
- `gross`: Печалба;
- `movie_title`: Име на филма;
- `num_voted_users`: Брой гласували потребители;
- `num_user_for_reviews`: Брой коментари на потребители;
- `language`: Език;
- `country`: Страна;
- `budget`: Бюджет на филма ;
- `title_year`: Година на излизане на филма;
- `imdb_score`: Оценка в imdb;
- `movie_facebook_likes`: Брой харесвания във facebook за филма ;

**Задача 1:**

а) Прочетете данните и ги запишете в data frame в `R`;

б) Генерирайте си подизвадка от 3000 наблюдения. За целта нека `f_nr` е вашият факултетен номер. Задайте състояние на генератора на случайни числа в `R` чрез `set.seed(f_nr)`. С помощта на *подходяща функция* генерирайте извадка *без връщане* на числата от 1 до 3800 като не забравяте да я запишете във вектор. Използвайте вектора, за да зашишете само редовете със съответните индекси в нов дейтафрейм и работете с него оттук нататък;

в) Изкарайте на екрана имената на променливите;

г) Направете summary на наблюдаваните величини;

д) Какъв вид данни (качествени/количествени, непрекъснати/дискретни) са записани във всяка от променливите?

е) Коя година за последен път е наблюдаван цветен филм?

ж) Изведете само по един път всяка от годините, през които е регистриран черно бял филм?

з) Постройте хистограма на разпределението на филмите според тяхната продължителност(duration) и анализирайте графиката.

и) Постройте кръстосана таблица и barplot на разпределението на филмите, снимани Germany, UK, Canada според това дали са цветни или черно бели.

к) Изследвайте съвместно `duration` и `color` с подходящ графичен метод. Забелязвате ли outlier-и? Сравнете извадковите средни и медианите в двете групи и направете извод;

**Задача 2:** За филмите с бюджет над 100 000 000 изследвайте съвместно променливите `num_voted_users` и `num_user_for_reviews` с подходящ графичен метод. Бихте ли казали, че съществува линейна връзка между тях? Намерете корелацията между величините и коментирайте стойността ѝ. Начертайте регресионна права (линейната функция, която най-добре приближава функционалната зависимост). Ако е наблюдаван нов филм, за който са гласували 700000 потребители (num_voted_users = 700000), какъв е очакваният брой потребители, които ще го коментират (num_user_for_reviews) на базата на линейния модел?

**Задача 3:** Нека $ABCD$ са последните 4 цифри на факултетния Ви номер. С помощта на `R` пресметнете:

- $\mathbb{P}(\min\{1, D\} < X \leq D + 3)$ за сл. в. $X \sim Bi(10, (A + B + C + D)/50)$;
- Стойността $z^*$, така че $\mathbb{P}(-z^* < Z \leq z^*) = (A + B + C + D)/50$ за сл. в. $Z \sim N(0, 1)$.

**Инструкции за предаване на контролната работа:**

- Предаването на контролното ще бъде през страницата на курса в moodle. В съответна факултативна група.
- Имате право да използвате само "Simple R"– John Verzani и документация. Като в началото ще имате 5 мин. да си изтеглите учебника и данните.
- Файлът който предавате в края на контролното трябва да е с име факултетният ви номер и да е в .R формат. Пример: 12345.R
- Прилага се кодът на `R`, както коментари и интерпретация на получените статистически резултати;
- Време за работа: 1 час.