Задача 4
Вариант 1

от Татаринова Никиты Алексеевича

к 22.04.2021

Условие

Файл "youtube_1.csv" содержит следующие сведения о видеороликах на YouTube (100 роликов):

- **n** – номер наблюдения;
- **id** – идентификатор ролика;
- **framerate** – число кадров в секунду;
- **frames** – общее число кадров в видео;
- **bitrate** – битрейт, Кбит/сек;
- **duration** – продолжительность ролика, сек;
- **size** – размер видеофайла, байт.

Для признаков **framerate**, **frames**, **bitrate**, **duration** и **size** рассчитайте две корреляционные матрицы – на основании коэффициентов Пирсона и Спирмена. Оцените значимость каждого коэффициента (проверьте гипотезу об отсутствии корреляции) и представьте полученные результаты в виде таблицы:

Коэффициент корреляции Пирсона

	framerate	**frames**	**bitrate**	**duration**	**size**
framerate	1	0.08	-0.02	0.04	0.02
frames	0.08	1	0.12	0.45**	0.29*
bitrate	-0.02	0.12	1	-0.03	0.72***
duration	0.04	0.45**	-0.03	1	0.36**
size	0.02	0.29*	0.72***	0.36**	1

- * – коэффициент значим на уровне 5%;
- ** – коэффициент значим на уровне 1%;
- *** – коэффициент значим на уровне 0.1%.

Коэффициенты, не отмеченные звёздочками, незначимы (нет оснований отвергнуть гипотезу об отсутствии корреляции на уровне 5%).

Сравните коэффициенты Пирсона и Спирмена, обратите внимание на случаи, когда два этих коэффициента существенно расходятся, если такие есть. Что такое "существенно", решайте сами. В случае существенного расхождения постройте диаграммы разбросы для тех пар признаков, тесноту связи между которыми коэффициенты измеряют по-разному, и попытайтесь объяснить причину расхождения.

Решение

В качестве языка программирования для решения данной задачи используется C++ (исходный код представлены в виде файла ".cpp"; скриншоты исходного кода в приложении).

Для начала, считаем данные, игнорируя номера наблюдений и идентификаторы роликов, так как для нашего анализа они не играют роли.

[В методе get_data читаем необходимые для анализа данные из файла "youtube_1.csv", сохраняя их в соответствующие массивы.]

Далее, для каждой пары рассматриваемых переменных вычисляем выборочные коэффициенты корреляции Пирсона и значения t-статистик.

$$r_{X,Y} = \widehat{Corr}(X,Y) = \frac{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)\cdot\left(Y_i-\overline{Y}\right)}{\sqrt{\sum\limits_{i=1}^{n}\left(X_i-\overline{X}\right)^2\times\sum\limits_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}}$$

$$t = \frac{r_{X,Y}\cdot\sqrt{n-2}}{\sqrt{1-r_{X,Y}^2}} \overset{H_0}{\sim} t_{n-2}$$

$$\begin{cases} H_0\colon\ \rho=0 \quad (X,Y \text{ независимы}) \\ H_A\colon\ \rho\neq 0 \quad\ \ (X,Y \text{ зависимы}) \end{cases}$$

[Для каждой пары рассматриваемых переменных вызываем метод test_zero_correlation_hypothesis_with_Pearson, в котором вычисляется $r_{X,Y}$ и значение t-статистики и сохранений этих значений в файл "Pearson_correlation.txt".]

Полученные значение коэффициентов корреляции Пирсона и будут значениями нашей таблицы. Осталось только проверить их значимость, для чего и нужно значение t-статистики:

- если $t < t_{98,\frac{0.05}{2}} = 1.984467404$, то нет оснований отвергнуть нулевую гипотезу и связь не выявлена (0 звёздочек);
- если $1.984467404 = t_{98,\frac{0.05}{2}} \leqslant t < t_{98,\frac{0.01}{2}} = 2.626931008$, то основная гипотеза отвергается в пользу альтернативной на уровне значимости 0.05 (1 звёздочка);
- если $2.626931008 = t_{98,\frac{0.01}{2}} \leqslant t < t_{98,\frac{0.001}{2}} = 3.39258811$, то основная гипотеза отвергается в пользу альтернативной на уровне значимости 0.01 (2 звёздочки);
- если $3.39258811 = t_{98,\frac{0.001}{2}} \leqslant t$, то основная гипотеза отвергается в пользу альтернативной на уровне значимости 0.001 (3 звёздочки).

На основании файла "Pearson_correlation.txt" получаем таблицу.

Коэффициент корреляции Пирсона

	framerate	**frames**	**bitrate**	**duration**	**size**
framerate	1	0.27**	0.24*	0.09	0.19
frames	0.27**	1	0.18	0.95***	0.87***
bitrate	0.24*	0.18	1	0.12	0.46***
duration	0.09	0.95***	0.12	1	0.78***
size	0.19	0.87***	0.46***	0.78***	1

Теперь, для каждой переменной получим соответствующий массив рангов. Для каждой пары рассматриваемых массивов рангов вычисляем коэффициенты ранговых корреляций Спирмана и значения t-статистик (воспользуемся приближением).

$$r_{X,Y}^S = r_{rank(X),rank(Y)}$$

$$t = \frac{r_{X,Y}^S\cdot\sqrt{n-2}}{\sqrt{1-\left(r_{X,Y}^S\right)^2}} \overset{H_0}{\sim} t_{n-2}$$

[Для каждой пары рассматриваемых массивов рангов вызываем метод test_zero_correlation_hypothesis_with_Spearman, в котором вычисляется $r^S_{X,Y}$ и значение t-статистики и сохранений этих значений в файл "Spearman_correlation.txt".]

В таком случае, условия составления таблицы остаются теми же. На основании файла "Spearman_correlation.txt" получаем таблицу.

Коэффициент корреляции Спирмена

	framerate	**frames**	**bitrate**	**duration**	**size**
framerate	1	0.37***	0.43***	0.11	0.38***
frames	0.37***	1	0.17	0.94***	0.67***
bitrate	0.43***	0.17	1	0.02	0.78***
duration	0.11	0.94***	0.02	1	0.58***
size	0.38***	0.67***	0.78***	0.58***	1

Проанализируем разницу для пар framerate – size и bitrate – size.

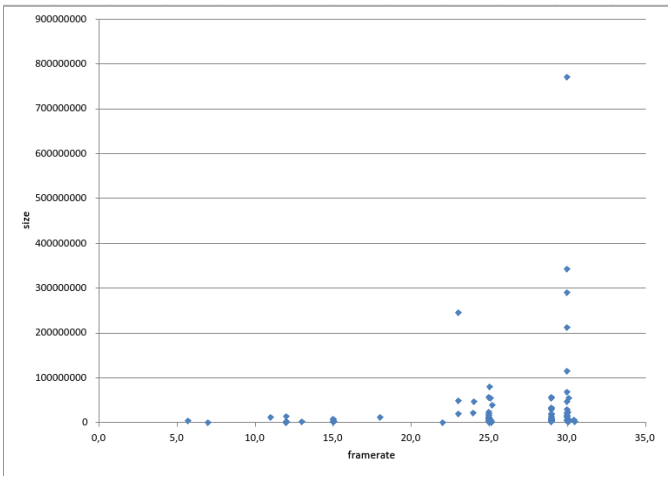

График framerate – size

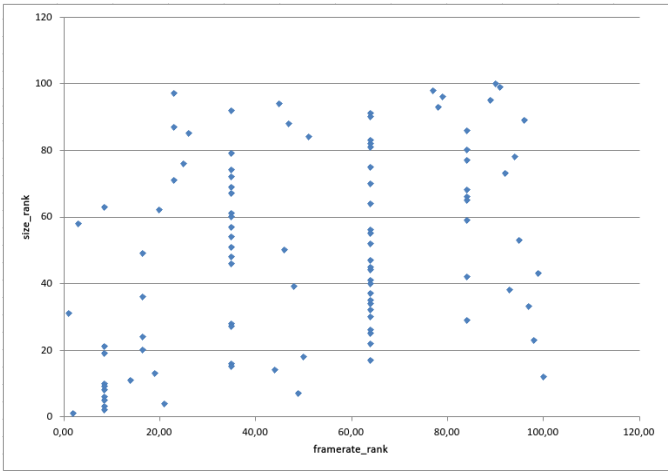

График framerate_rank – size_rank

График framerate-size больше монотонен, чем линеен. График framerate_rank – size_rank рассеян, но более линеен.

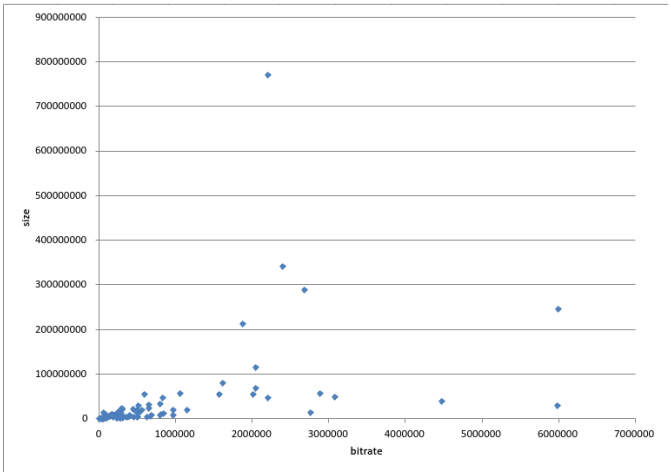

График bitrate – size

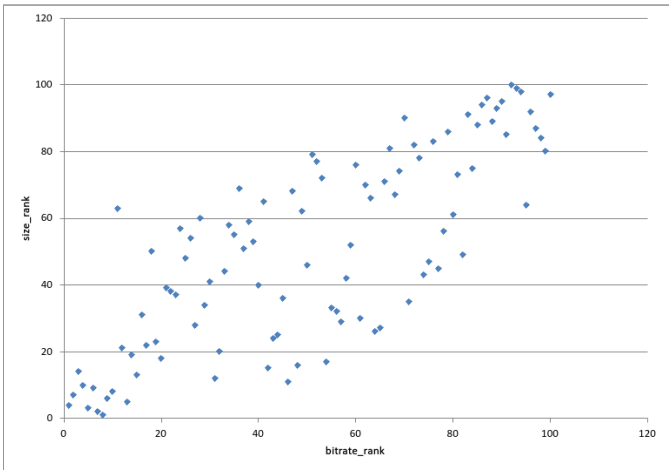

График bitrate_rank – size_rank

График bitrate – size выглядит достаточно линейным, за исключением нескольких точек. График bitrate_rank – size_rank относительно рассеян.