Домашнее задание 4 Вариант 4

Выполнил: Мосолков Евгений Николаевич БПИ196

Примечание:

Вычисления были сделаны при помощи Python и Jupiter notebook. Среди использованных библиотек – pandas, scipy, matplotlib

Задание:

1. 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
2. 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 разброса для тех пар признаков, тесноту связи между которыми коэффициенты измеряют по-разному, и попытайтесь объяснить причину расхождения. Если вы не видите никаких существенных расхождений между двумя матрицами, просто постройте диаграмму рассеяния для случая, где разность коэффициентов Пирсона и Спирмена наибольшая

Решение:

1. Найдем матрицу корреляции Пирсона

Считаем сумму средних значений столбцов, затем среднее арифметическое столбов. Теперь считаем отклонения от среднего арифметического для каждого значения из каждого столбца. Возводим отклонения в квадрат и считаем сумму отклонений.

Далее считаем коэффициенты корреляции по формуле:

$$r_{xy}=\frac{\sum(x¿¿i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum(x¿¿i-\overline{x})^2*\sum(y_i-\overline{y})^2¿}¿}$$

Определяем значимость коэффициента корреляции по формуле:

$$t=r_{xy}\frac{\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$$

Расставляем метки, сравнивая значения со значениями из таблицы Стьюдента и получаем таблицу:

	framerate	frames	bitrate	duration	size
framerate	1	0.300**	0.343***	0.090	0.290**
frames	0.300**	1	0.009	0.902***	0.549***
bitrate	0.343***	0.009	1	-0.041	0.586***
duration	0.090	0.902***	-0.041	1	0.443***
size	0.290**	0.549***	0.586***	0.443***	1

2. Найдем матрицу корреляции Спримена

Для этого необходимо про ранжировать данные по столбцам, затем вычислить разность рангов по каждому случаю. Полученную разность нужно возвести в квадрат. Далее находим сумму квадратов разностей.

Теперь вычислим значения, которые попадут в матрицу по формуле:

$$p=1-6\frac{\Sigma d^2}{n^3-n}$$

Определяем значимость коэффициента корреляции по формуле:

$$t=p\frac{\sqrt{n-2}}{\sqrt{1-p^2}}$$

Расставляем метки, сравнивая значения со значениями из таблицы Стьюдента и получаем таблицу:

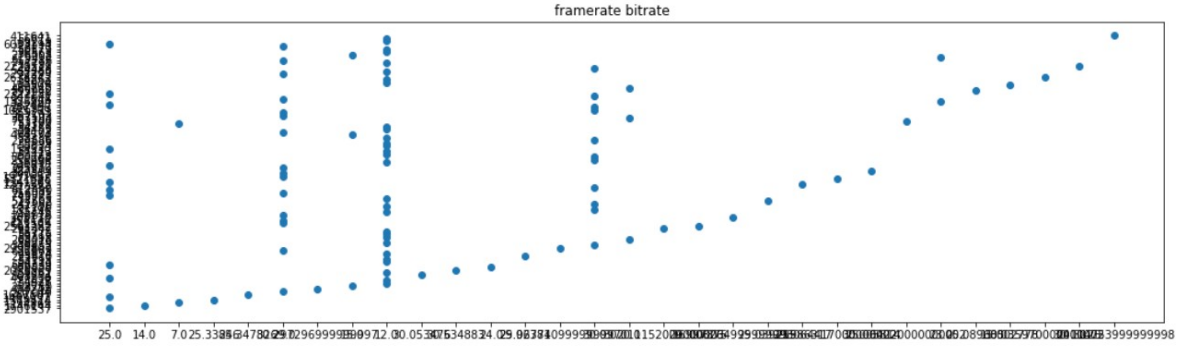
	framerate	frames	bitrate	duration	size
framerate	1	0.510***	0.648***	0.225*	0.605***
frames	0.510***	1	0.373***	0.937***	0.799***
bitrate	0.648***	0.373***	1	0.152	0.807***
duration	0.225*	0.937***	0.152	1	0.653***
size	0.605***	0.799***	0.807***	0.653***	1

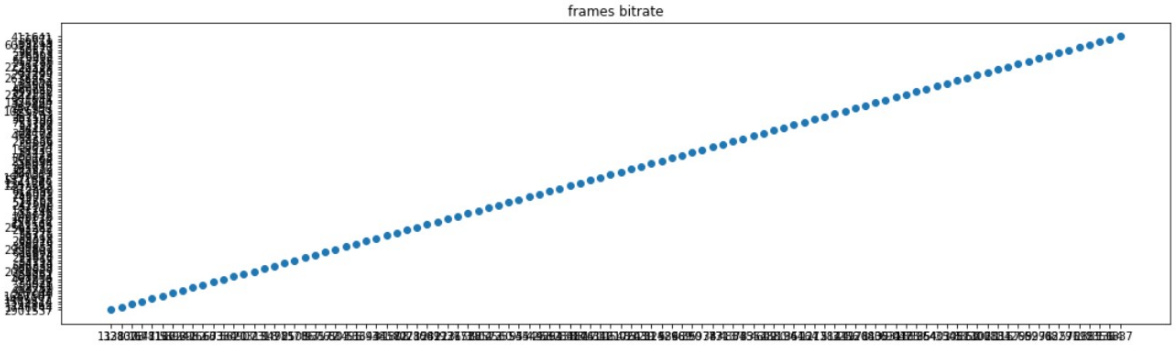
	framerate	frames	bitrate	duration	size
framerate	1	0.300**	0.343***	0.090	0.290**
frames	0.300**	1	0.009	0.902***	0.549***
bitrate	0.343***	0.009	1	-0.041	0.586***
duration	0.090	0.902***	-0.041	1	0.443***
size	0.290**	0.549***	0.586***	0.443***	1

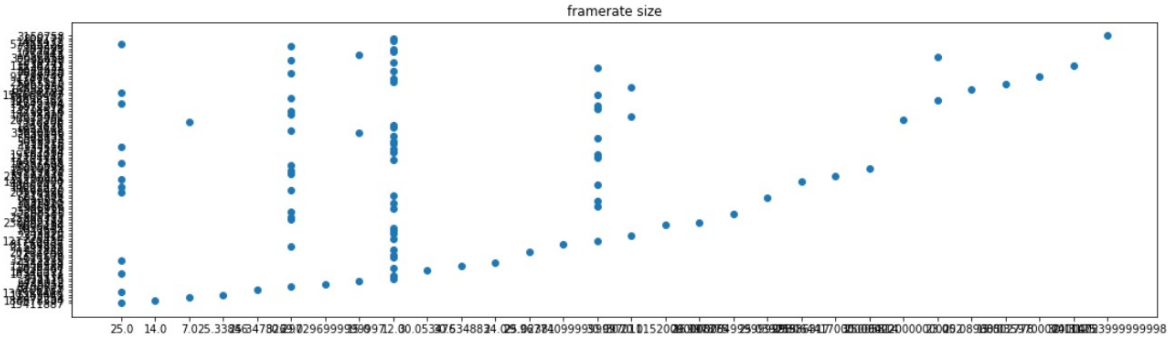
Наибольший разброс наблюдается у следующих пар параметров:

1. framerate и bitrate
2. frames и bitrate
3. framerate и size

Построим диаграммы для каждой из пар параметров:


framerate bitrate


frames bitrate


framerate size