МГТУ им. Н. Э. Баумана, кафедра ИУ5

курс “Технологии машинного обучения”

Лабораторная работа №1

«Разведочный анализ данных. Исследование и визуализация данных»

ВЫПОЛНИЛ:

Пученков Д.О.

Группа: ИУ5-61Б

ПРОВЕРИЛ:

Гапанюк Ю.Е.

Москва, 2020 г.

Цель лабораторной работы: изучение различных методов визуализации данных.

Задание:

- Выбрать набор данных (датасет).

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

- Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

- Сформировать отчет и разместить его в своем репозитории на github.

Выполненная работа:

Данный датасет содержит информацию по всем видеозаписям, попавшим в тренды youtube в Соединенных Штатах:

1. **video_id** – id видеозаписи
2. **trending_date** – дата попадания в тренды
3. **title** - заголовок
4. **channel_title** – название канала
5. **category_id** – категория видеозаписи
6. **publish_time** – дата публикации
7. **tags** - тэги
8. **views** – количество просмотров
9. **likes** – количество лайков
10. **dislikes** – количество дислайков
11. **comment_count** – количество комментариев
12. **thumbnail_link** – ссылка на видеозапись
13. **comments_disabled** – отключены ли комментарии
14. **ratings_disabled** – отключены ли лайки
15. **video_error_or_removed** – статус видеозаписи
16. **description** - описание

Текст программы и экранные формы с примерами выполнения программы:

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
sns.set(style="ticks")
```

```
data = pd.read_csv('../datasets/USvideos.csv', sep=",")
```

```
data.head()
```

	video_id	trending_date	title	channel_title	category_id	publish_time	tags
0	2kyS6SvSYSE	17.14.11	WE WANT TO TALK ABOUT OUR MARRIAGE	CaseyNeistat	22	2017-11-13T17:13:01.000Z	SHANtell martin
1	1ZAPwfrtAFY	17.14.11	The Trump Presidency: Last Week Tonight with J...	LastWeekTonight	24	2017-11-13T07:30:00.000Z	last week tonight trump presi week ...
2	5qpjK5DgCt4	17.14.11	Racist Superman | Rudy Mancuso, King Bach & Le...	Rudy Mancuso	23	2017-11-12T19:05:24.000Z	racist superman|"rudy"|"mancuso"|"
3	puqaWrEC7tY	17.14.11	Nickelback Lyrics: Real or Fake?	Good Mythical Morning	24	2017-11-13T11:00:04.000Z	rhett and link|"gmm"|"good m morning"|"...
4	d380meD0W0M	17.14.11	I Dare You: GOING BALDI?	nigahiga	24	2017-11-12T18:01:41.000Z	ryan|"higa"|"higatv"|"nigahiga you"|"...

```
data.shape
```

```
(40949, 16)
```

```
data.columns
```

```
Index(['video_id', 'trending_date', 'title', 'channel_title', 'category_id',
       'publish_time', 'tags', 'views', 'likes', 'dislikes', 'comment_count',
       'thumbnail_link', 'comments_disabled', 'ratings_disabled',
       'video_error_or_removed', 'description'],
      dtype='object')
```

```
data.dtypes
```

```
video_id                 object
trending_date            object
title                    object
channel_title            object
category_id               int64
publish_time             object
tags                     object
views                     int64
likes                     int64
dislikes                  int64
comment_count             int64
thumbnail_link           object
comments_disabled          bool
ratings_disabled           bool
video_error_or_removed     bool
description              object
dtype: object
```

```
for col in data.columns:
    temp_null_count = data[data[col].isnull()].shape[0]
    print('{}- {}'.format(col,temp_null_count))
```

```
video_id- 0
trending_date- 0
title- 0
channel_title- 0
category_id- 0
publish_time- 0
tags- 0
views- 0
likes- 0
dislikes- 0
comment_count- 0
thumbnail_link- 0
comments_disabled- 0
ratings_disabled- 0
video_error_or_removed- 0
description- 570
```

```
data.describe()
```

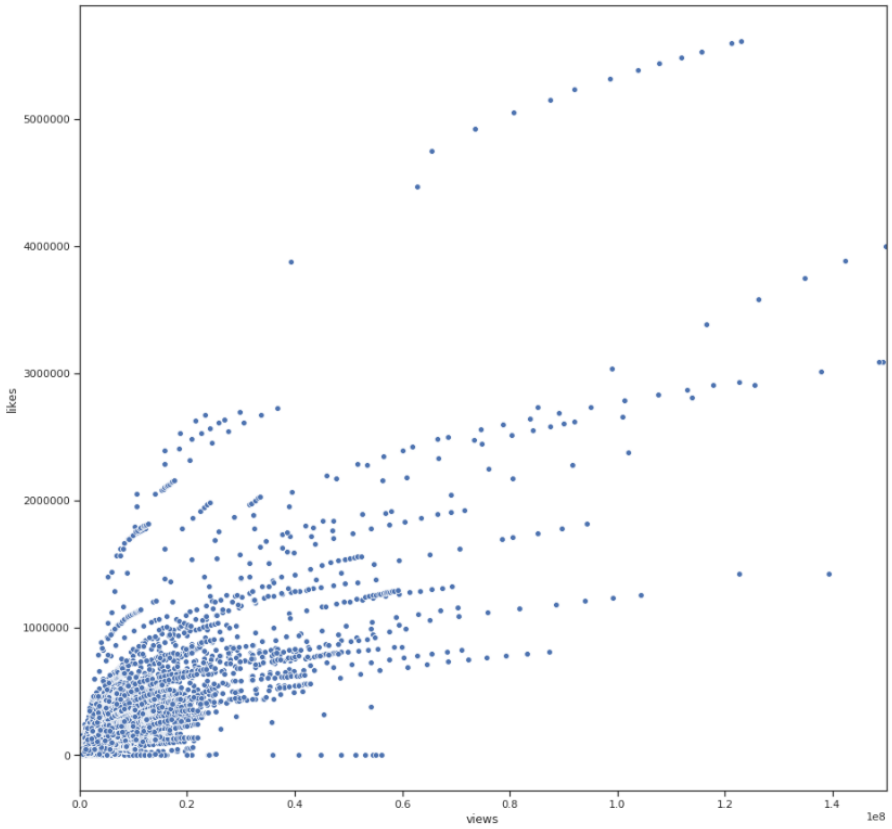
	category_id	views	likes	dislikes	comment_count
count	40949.000000	4.094900e+04	4.094900e+04	4.094900e+04	4.094900e+04
mean	19.972429	2.360785e+06	7.426670e+04	3.711401e+03	8.446804e+03
std	7.568327	7.394114e+06	2.288853e+05	2.902971e+04	3.743049e+04
min	1.000000	5.490000e+02	0.000000e+00	0.000000e+00	0.000000e+00
25%	17.000000	2.423290e+05	5.424000e+03	2.020000e+02	6.140000e+02
50%	24.000000	6.818610e+05	1.809100e+04	6.310000e+02	1.856000e+03
75%	25.000000	1.823157e+06	5.541700e+04	1.938000e+03	5.755000e+03
max	43.000000	2.252119e+08	5.613827e+06	1.674420e+06	1.361580e+06

Визуальное исследование датасета.

Визуально исследовать наш датасет мы будем при помощи диаграмм рассеивания и гистограмм. С помощью диаграммы рассеивания мы сможем оценить существуют ли отношения или корреляция между этими двумя переменными, например, для максимальных и минимальных значений.

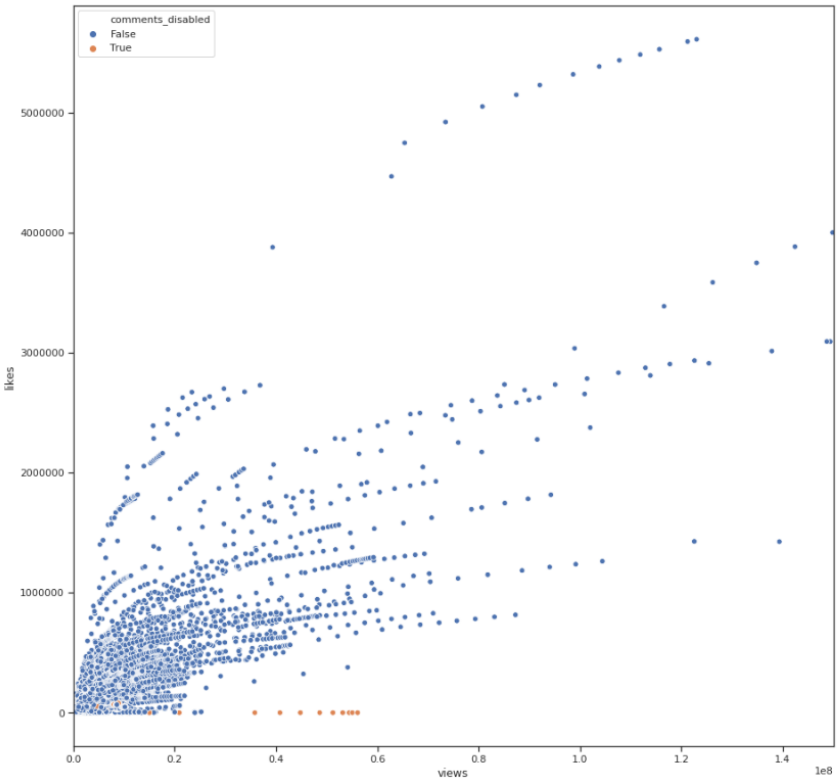
```
fig, ax = plt.subplots(figsize=(15,15))
ax.set(xlim=(0, 1.5*10**8))
sns.scatterplot(ax=ax, x='views',y='likes',data=data)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f07d0825c90>
```



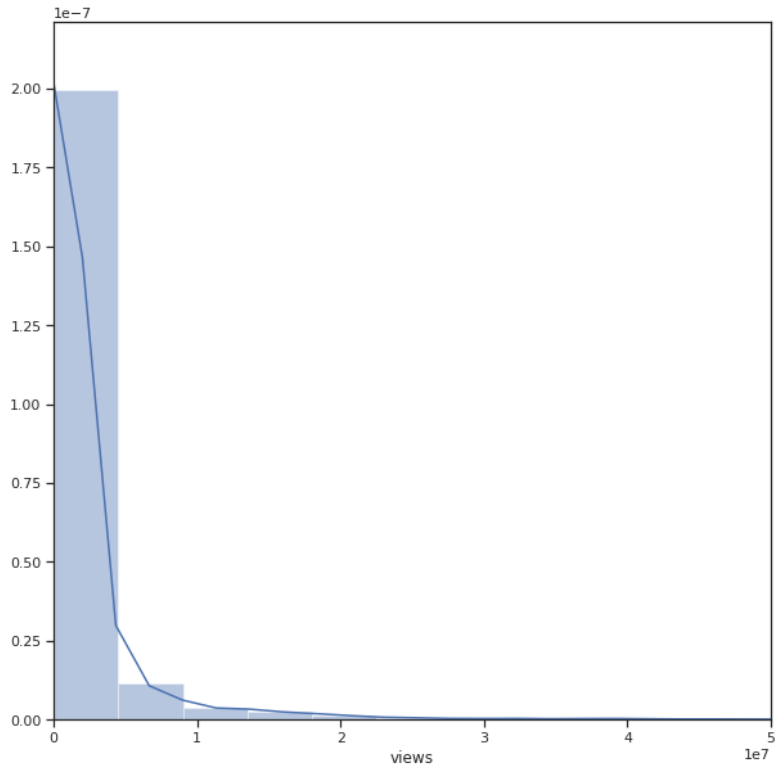

```
fig, ax = plt.subplots(figsize=(15,15))
ax.set(xlim=(0,1.5*10**8))
sns.scatterplot(ax=ax, x='views', y='likes', data=data, hue='comments_disabled')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f07d0826690>
```



```
fig,ax = plt.subplots(figsize=(10,10))
ax.set(xlim=(0,0.5*10**8))
sns.distplot(data['views'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f07d078f690>
```

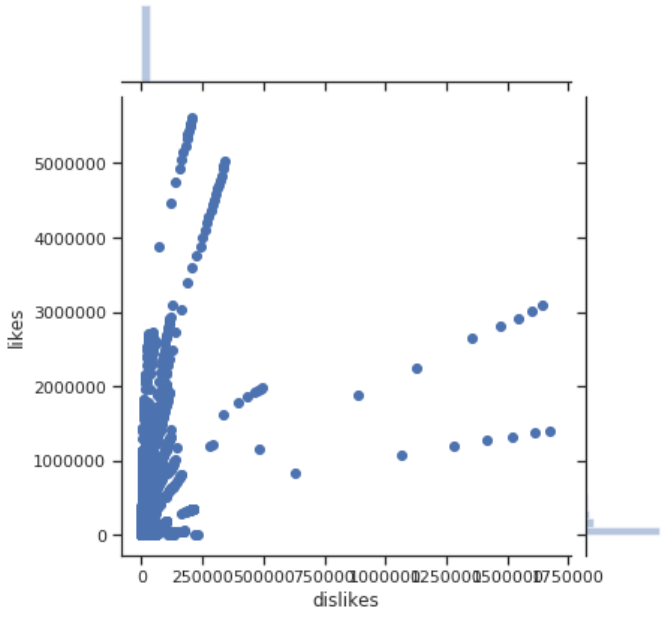


С помощью гистограммы мы можем оценить плотность вероятности распределения данных для поля “views ”.

Также для наглядности можно построить jointplot - комбинацию гистограмм и диаграмм рассеивания.

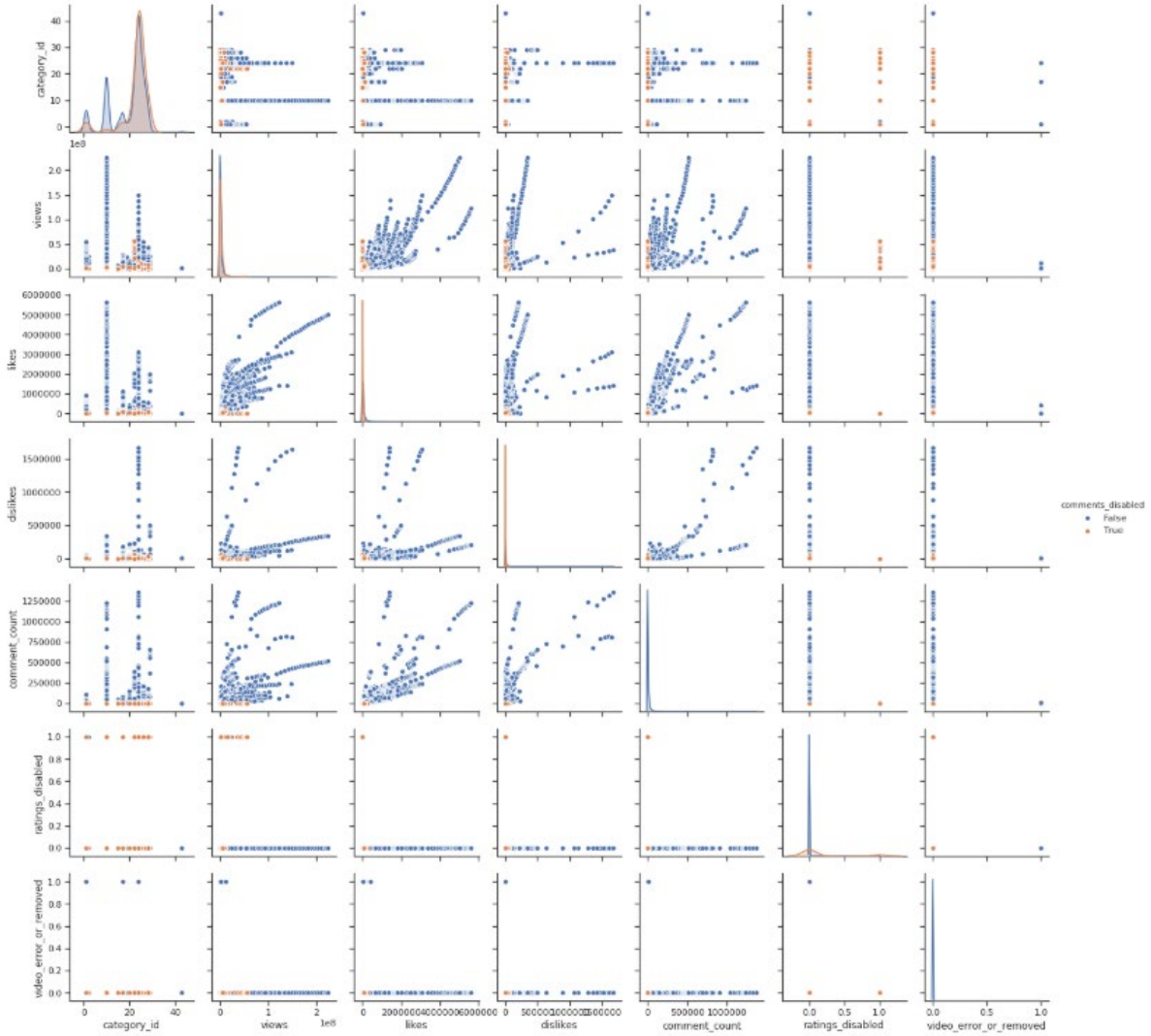
```
sns.jointplot(x='dislikes', y='likes', data=data)
```

```
<seaborn.axisgrid.JointGrid at 0x7f07cb6e5610>
```



```
sns.pairplot(data, hue="comments_disabled")
```

```
<seaborn.axisgrid.PairGrid at 0x7fe49575d750>
```



Информация о корреляции признаков

Проверка корреляции признаков позволяет решить две задачи:

1. Понять какие признаки (колонки набора данных) наиболее сильно коррелируют с целевым признаком. Именно эти признаки будут наиболее информативными для моделей машинного обучения. Признаки, которые слабо коррелируют с целевым признаком, можно попробовать исключить из построения модели, иногда это повышает качество модели. Нужно отметить, что некоторые алгоритмы машинного обучения автоматически определяют ценность того или иного признака для построения модели.
2. Понять какие нецелевые признаки линейно зависимы между собой. Линейно зависимые признаки, как правило, очень плохо влияют на качество моделей. Поэтому если несколько признаков линейно зависимы, то для построения модели из них выбирают какой-то один признак.

```
data.corr()
```

	category_id	views	likes	dislikes	comment_count	comments_disabled	ratings_disa
category_id	1.000000	-0.168231	-0.173921	-0.033547	-0.076307	0.048949	-0.013506
views	-0.168231	1.000000	0.849177	0.472213	0.617621	0.002677	0.015355
likes	-0.173921	0.849177	1.000000	0.447186	0.803057	-0.028918	-0.020888
dislikes	-0.033547	0.472213	0.447186	1.000000	0.700184	-0.004431	-0.008230
comment_count	-0.076307	0.617621	0.803057	0.700184	1.000000	-0.028277	-0.013819
comments_disabled	0.048949	0.002677	-0.028918	-0.004431	-0.028277	1.000000	0.319230
ratings_disabled	-0.013506	0.015355	-0.020888	-0.008230	-0.013819	0.319230	1.000000
video_error_or_removed	-0.030011	-0.002256	-0.002641	-0.001853	-0.003725	-0.002970	-0.001526

```
data.corr(method='pearson')
```

	category_id	views	likes	dislikes	comment_count	comments_disabled	ratings_disa
category_id	1.000000	-0.168231	-0.173921	-0.033547	-0.076307	0.048949	-0.013506
views	-0.168231	1.000000	0.849177	0.472213	0.617621	0.002677	0.015355
likes	-0.173921	0.849177	1.000000	0.447186	0.803057	-0.028918	-0.020888
dislikes	-0.033547	0.472213	0.447186	1.000000	0.700184	-0.004431	-0.008230
comment_count	-0.076307	0.617621	0.803057	0.700184	1.000000	-0.028277	-0.013819
comments_disabled	0.048949	0.002677	-0.028918	-0.004431	-0.028277	1.000000	0.319230
ratings_disabled	-0.013506	0.015355	-0.020888	-0.008230	-0.013819	0.319230	1.000000
video_error_or_removed	-0.030011	-0.002256	-0.002641	-0.001853	-0.003725	-0.002970	-0.001526

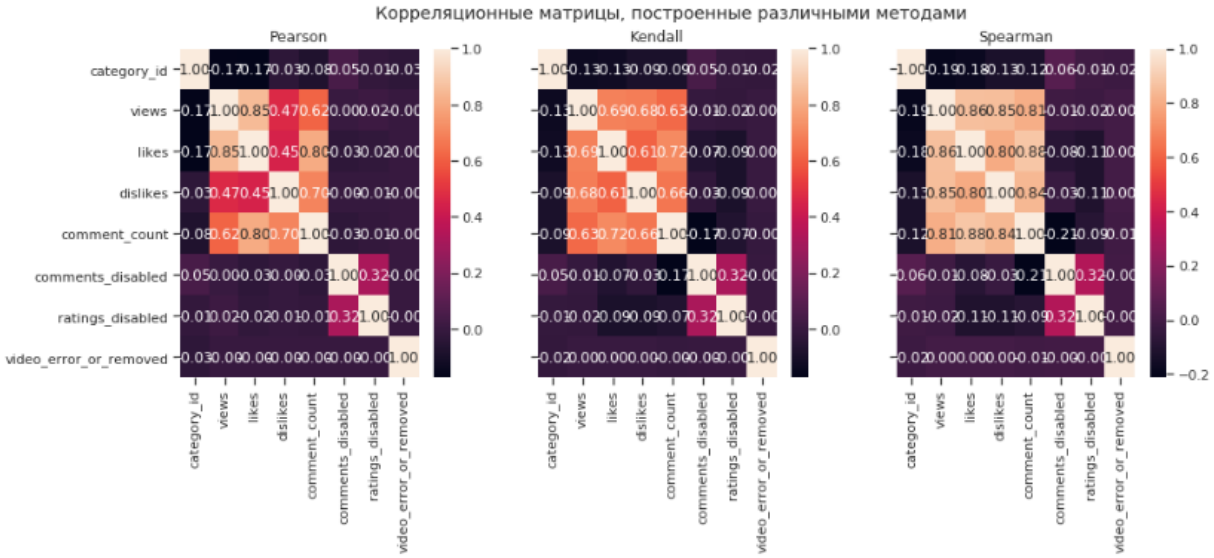
```
data.corr(method='kendall')
```

	category_id	views	likes	dislikes	comment_count	comments_disabled	ratings_disa
category_id	1.000000	-0.134824	-0.126356	-0.089573	-0.087364	0.050209	-0.010445
views	-0.134824	1.000000	0.691329	0.680772	0.632967	-0.007929	-0.015963
likes	-0.126356	0.691329	1.000000	0.614065	0.715186	-0.065620	-0.090661
dislikes	-0.089573	0.680772	0.614065	1.000000	0.662717	-0.026590	-0.090218
comment_count	-0.087364	0.632967	0.715186	0.662717	1.000000	-0.173964	-0.071621
comments_disabled	0.050209	-0.007929	-0.065620	-0.026590	-0.173964	1.000000	0.319230
ratings_disabled	-0.010445	-0.015963	-0.090661	-0.090218	-0.071621	0.319230	1.000000
video_error_or_removed	-0.017640	0.000752	0.001630	0.000611	-0.004823	-0.002970	-0.001526

```
data.corr(method='spearman')
```

	category_id	views	likes	dislikes	comment_count	comments_disabled	ratings_disa
category_id	1.000000	-0.190188	-0.179136	-0.126461	-0.124302	0.058190	-0.012105
views	-0.190188	1.000000	0.862553	0.854176	0.807619	-0.009711	-0.019551
likes	-0.179136	0.862553	1.000000	0.798874	0.878363	-0.080365	-0.111033
dislikes	-0.126461	0.854176	0.798874	1.000000	0.838240	-0.032555	-0.110459
comment_count	-0.124302	0.807619	0.878363	0.838240	1.000000	-0.213005	-0.087694
comments_disabled	0.058190	-0.009711	-0.080365	-0.032555	-0.213005	1.000000	0.319230
ratings_disabled	-0.012105	-0.019551	-0.111033	-0.110459	-0.087694	0.319230	1.000000
video_error_or_removed	-0.020444	0.000921	0.001996	0.000748	-0.005905	-0.002970	-0.001526

```
fig, ax = plt.subplots(1, 3, sharex='col', sharey='row', figsize=(15,5))
sns.heatmap(data.corr(method='pearson'), ax=ax[0], annot=True, fmt='.2f')
sns.heatmap(data.corr(method='kendall'), ax=ax[1], annot=True, fmt='.2f')
sns.heatmap(data.corr(method='spearman'), ax=ax[2], annot=True, fmt='.2f')
fig.suptitle('Корреляционные матрицы, построенные различными методами')
ax[0].title.set_text('Pearson')
ax[1].title.set_text('Kendall')
ax[2].title.set_text('Spearman')
```



Корреляционная матрица содержит коэффициенты корреляции между всеми парами признаков, она симметрична относительно главной диагонали. На главной диагонали расположены единицы (корреляция признака самого с собой). На основе корреляционной матрицы можно сделать выводы, которые помогут с решениями задач или для определения ненужных в выборке значений.