Университет ИТМО

Практическая работа №2

по дисциплине «Визуализация и моделирование»

Автор: Новожилова Анна Владимировна
Поток: ВИМ 1.2
Группа: К3222
Факультет: ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 2021г.

Ход работы

Описание шкал

На основе изученного материала, данные каждого столбца датасета были распределены по шкалам. Результаты представлены в таблице ниже.

Соответсвие столбцов и шкал

Название столбца	**Шкала**
mal id	Интервальная
aired from	Относительная
aired to	Относительная
duration	Номинальная
episodes	Относительная
genres	rating
score	Интервальная
source	Номинальная
status	Номинальная
studios	Номинальная
tittle	Номинальная
type	Номинальная

Результаты визуализации

В этой работе с помощью библиотеки Pandas данные были обработаны и представлены в виде простейших графиков и диаграмм.

Для начала решено было посчитать какое количество эпизодов содержит в себе каждое аниме, и какое количество эпизодов встречается чаще всего. С помощью математических функций было выявлено, что в среднем в одном аниме содержится десятть серий. Однако стандартное отклонение было довольно велико, поэтому возникла необходимость в подсчете частоты встречаемости значений.

```
episodes_data = [(episodes, df["episodes"].to_list().count(episodes))
                 for episodes in df["episodes"].unique()
                 if df["episodes"].to_list().count(episodes) > 8]
episodes_data = sorted(episodes_data, key=operator.itemgetter(1))
episodes_data
```

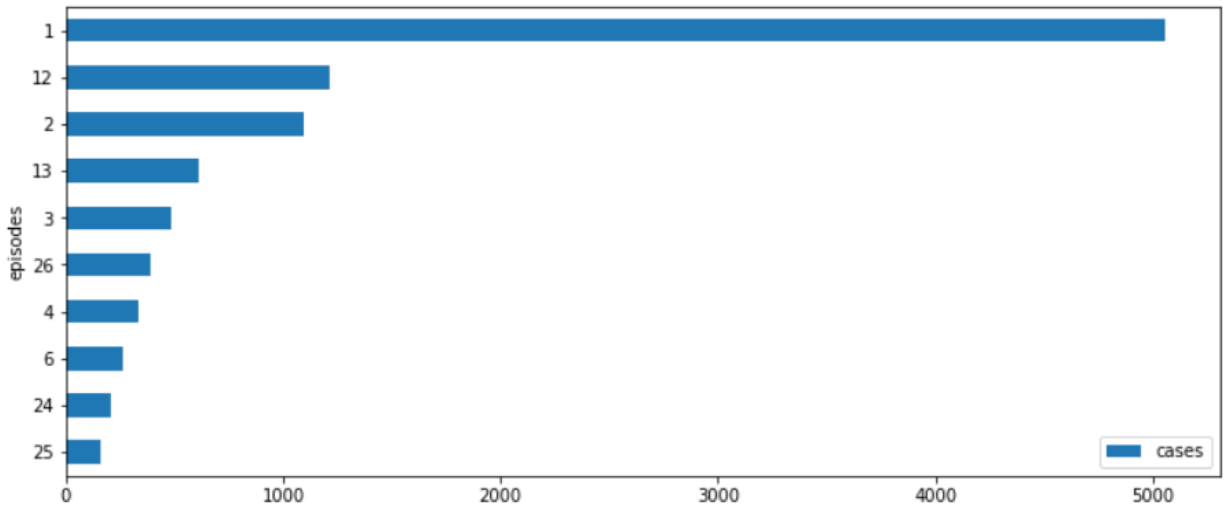
```
episodes_amount = []
cases = []
for categ, count in episodes_data:
  if count>150:
    episodes_amount.append(categ)
    cases.append(count)
```

```
inf_df = pd.DataFrame({"episodes": episodes_amount, "cases": cases})
fig, axes = plt.subplots(nrows=1, ncols=1)
inf_df.plot.barh(ax=axes, x="episodes", y="cases", figsize=(12, 5));
```

Обработка списка с количеством серий

Столбец с количеством эпизодов был переведен в список, отстортирован и подсчитана встречаемость каждого значения. Значения, которые встречаются достаточно редко (менее 150 раз) были отброшены, потому как в датасете 11 тысяч строк, и такие маленькие значения не характеризуют выборку.

В итоге по получившейся гистограмме видно, что большинство аниме в датасете - односерийные. Предположительно - фильмы.



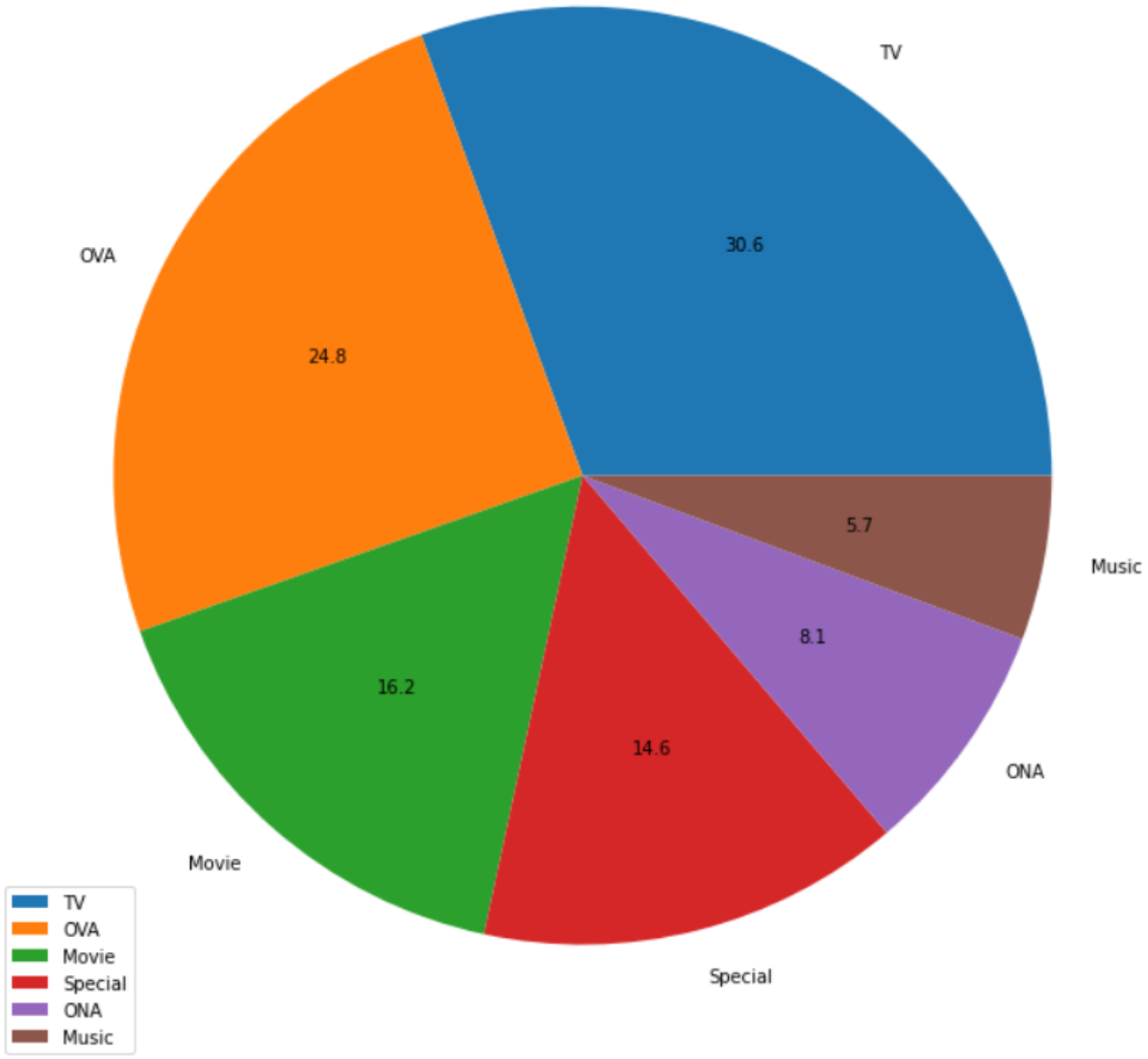
Гистограмма количества серий в аниме

Проверим гипотезу о том, что большая часть аниме в датасете - это полнометражные фильмы. Для этого переведем столбец type в список, и также

посчитаем количество одинаковых значений, сгруппируем их точно также, как с предыдущим столбцом. Для этих значчений наиболее наглядным графиком кажется круговая диаграмма.

```
types_info = pd.DataFrame({"": cases},
                          index=types)
fig, axes = plt.subplots(nrows=1, ncols=1)
types_info.plot.pie(ax=axes, y="",
                    autopct="%.1f",
                    fontsize=10,
                    figsize=(12, 12));
```

Построение диаграммы



Круговая диаграмма встречаемости типов аниме

Как ни странно, большинство аниме - это все-таки многосерийные аниме, которые выпускали в телеэфир. Однако значительную часть, почти такую же, как и сериалы, составляют ONA - спешлы к сезонам или к манге. Они также состоят в большинстве своем из 1 серии.

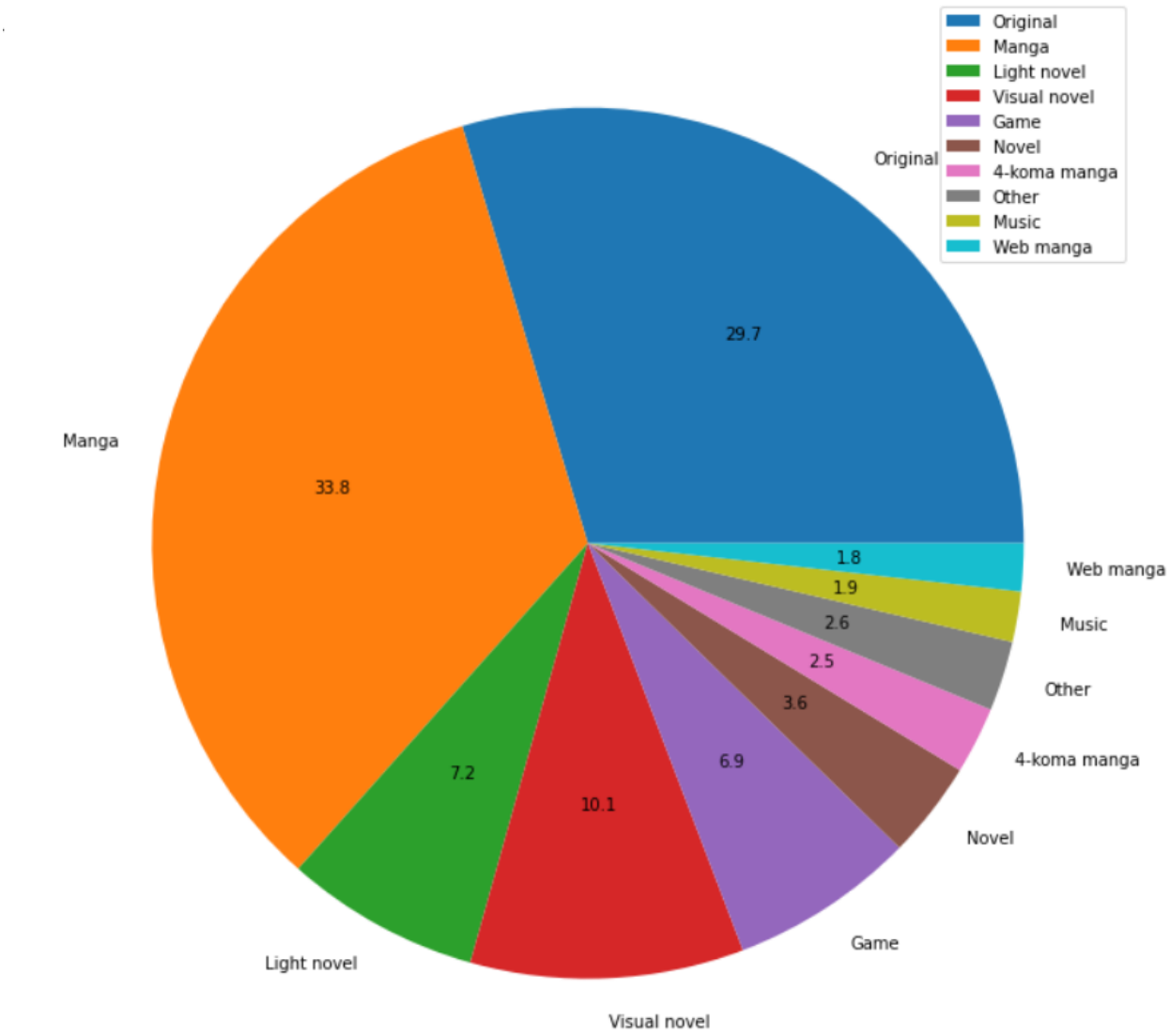
Дальше проверялось, из каких источников авторы берут сюжеты для своих аниме. Была выдвинута гипотеза, что большинство аниме имеет источником мангу или ранобэ.

```
source_data = {name: df["source"].to_list().count(name) for name in df["source"].unique()}
cases = []
sources = []
for source, cases_num in source_data.items():
  if source != "Unknown" and cases_num >150:
    sources.append(source)
    cases.append(cases_num)

source_info = pd.DataFrame({"": cases},
                        index=sources)
fig, axes = plt.subplots(nrows=1, ncols=1)
source_info.plot.pie(ax=axes, y="",
                 autopct="%.1f",
                 fontsize=10,
                 figsize=(12, 12));
```

Создание диаграммы источников

Маленькие значения были отброшены, потому что они не характеризуют выборку и только перегружают диаграмму.



Круговая диаграмма источников сюжета

Гипотеза подтвердилась лишь частично. Неожиданно, ранобэ оказались источником всего лишь 7 процентов аниме, но манга действительно играет большую роль в аниме индустрии. Очень большое количество аниме не имеет первоисточника.

Далее проверялось, какая есть тенденция в выставлении оценок, какие оценки ставят чаще всего. Для этого было решено использовать обыкновенный график, потому как шкала оценок непрерывна.

```
import matplotlib.pyplot as plt
```

```
score_df = pd.DataFrame.from_dict(data=score_data, orient="index", columns=["score"])
score_df.head()
```

```
score_df.plot();
```

Создание датасета и графика пользовательских оценок

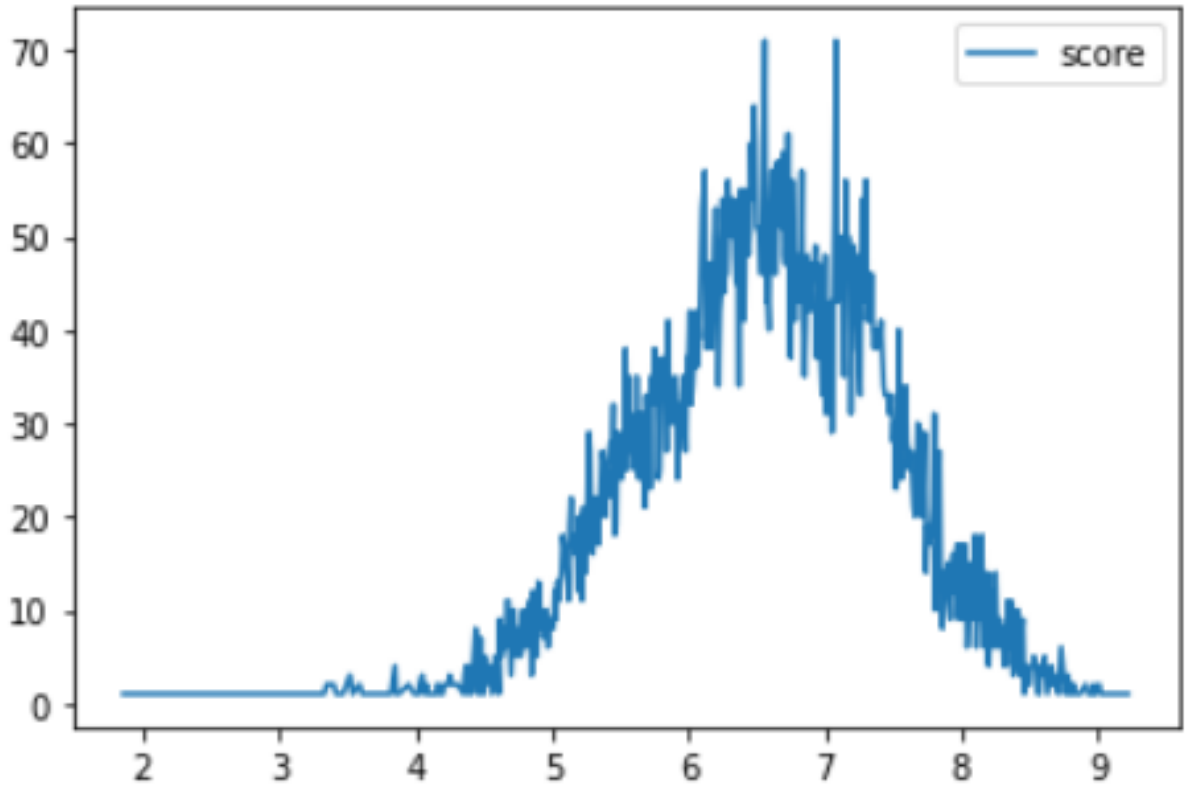


График пользовательских оценок

Как видно из графика, оценки почти подчиняются нормальному распределению и имеют пик в средних значениях 6-8 баллов.

Известно, что в датасете есть информация как о законченных аниме, так и о все еще продолжающих выходить в эфир. Посмотрим, насколько сильно разнятся оценки у вышедших и незаконченных аниме.

Для этого создадим новый датасет, в который будут входить только законченные аниме, и наложим его график на общий.

```
finished_df = df.loc[df["status"] == "Finished Airing"]
f_num = finished_df.shape[0]
f_percent = f_num / df.shape[0] * 100
print(f_num, f_percent, sep="\n")
finished_df.head()
```

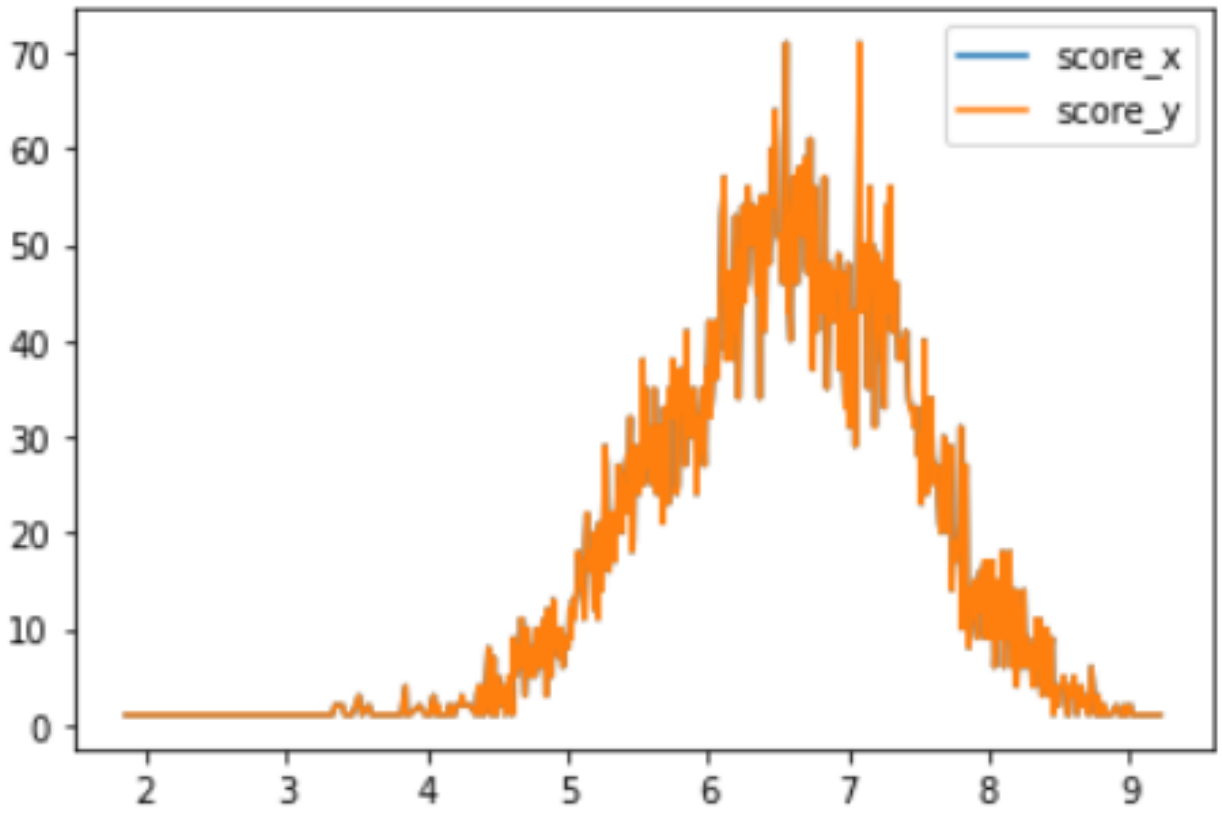
Создание базы данных с завершенными аниме

```
score_all = score_df.merge(score_finished_df, left_index=True, right_index=True)
score_all.head()
```

	score_x	score_y
1.86	1	1
2.04	1	1
2.25	1	1
2.33	1	1
2.35	1	1

```
score_all.plot();
```

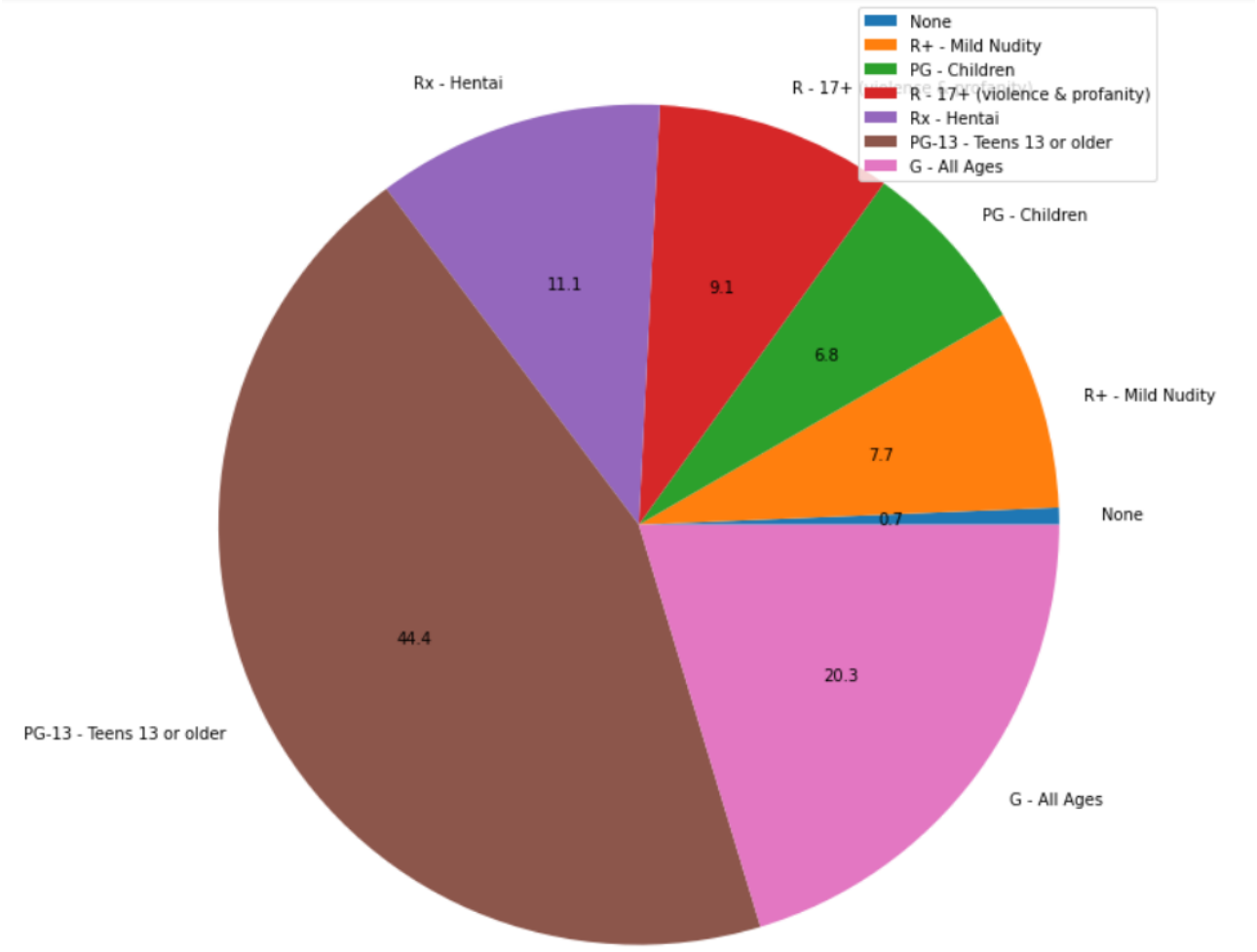
Объединение сведений об оценках



Объединенные графики

Данный график показал, что несмотря на наличие в датасете незавершенных аниме, они являются выбросами, так как оценки завершенных аниме почти идеально накладываются на общие оценки.

Уже описанным выше способом найдем распределение рейтинга в датасете.



Возрастной рейтинг аниме

Как видно из круговой диаграммы, большинство аниме созданы для аудитории от 13 лет. Это показывает, что анимеиндустрия во многом нацелена именно на подростковую аудиторию.

Далее рассматривался столбец с длительностью одной серии аниме. Также значения столбца были переведены в список, отсортированы и посчитано количество вхождений.

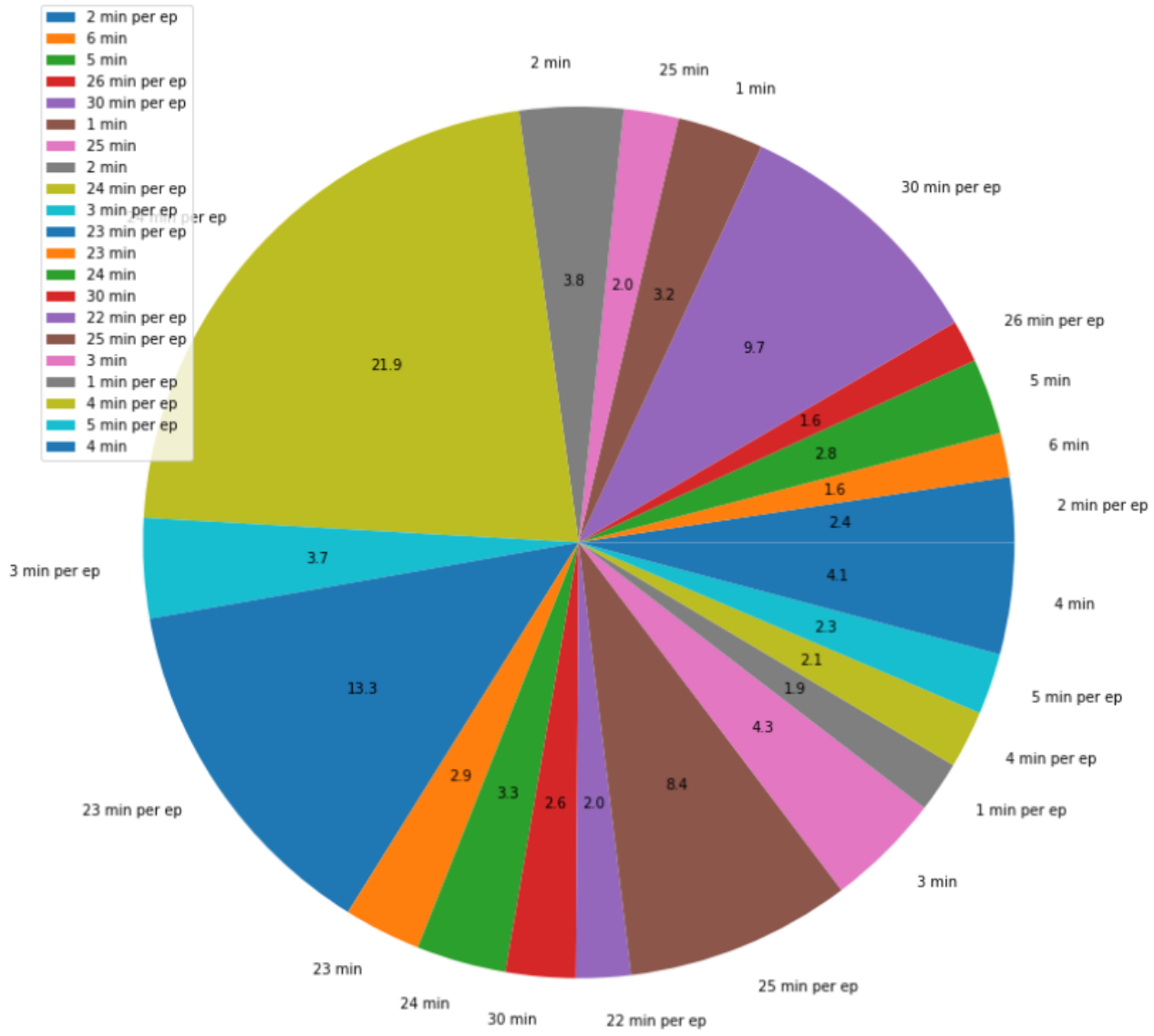


Диаграмма всех длин серий аниме

Как видно из рисунка, диаграмма получилась чиатемая, но слишком перегруженная и не наглядная. Однако маленькие значения составляют достаточно большую часть диаграммы, поэтому просто откидывать их было бы неразумно. Поэтому все значения с количеством вхождений меньше 500 были сгруппированы в категорию Other.

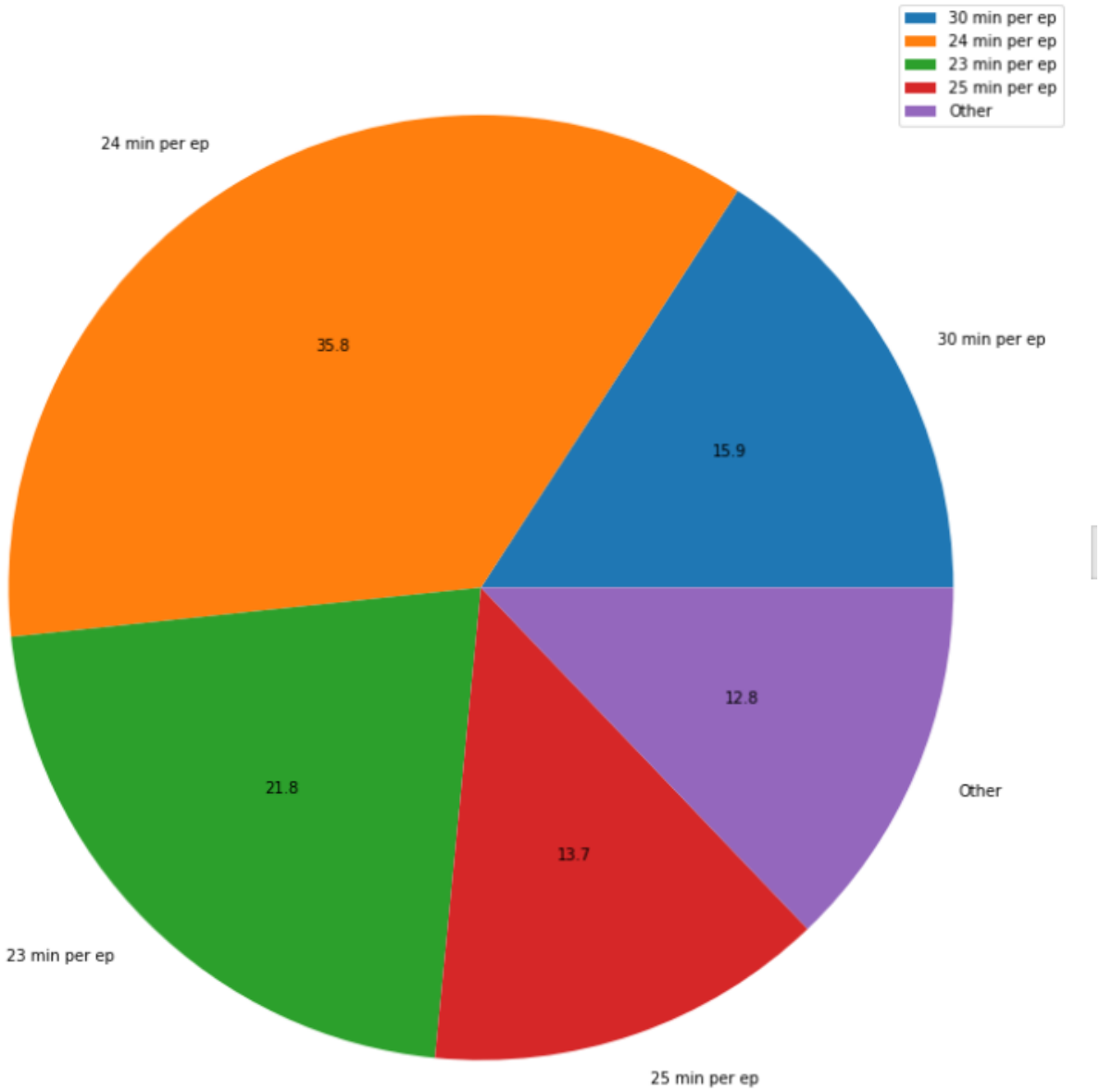


Диаграмма всех длин серий аниме

Теперь уже видно, что чаще всего встречаются значения, близкие к 24 минутам на серию. Это соответствует тому, что большая часть аниме - это сериалы и спэшлы.

Далее решено было оценить частоту встречаемости жанров в аниме. Это было проблематично, потому что значения этого столбца изначально были списком, но pandas преобразовал их в строку. Для того, чтобы получить значения, потребовалось разбить строку в ряды по разделителю, затем очистить все от лишних символов.

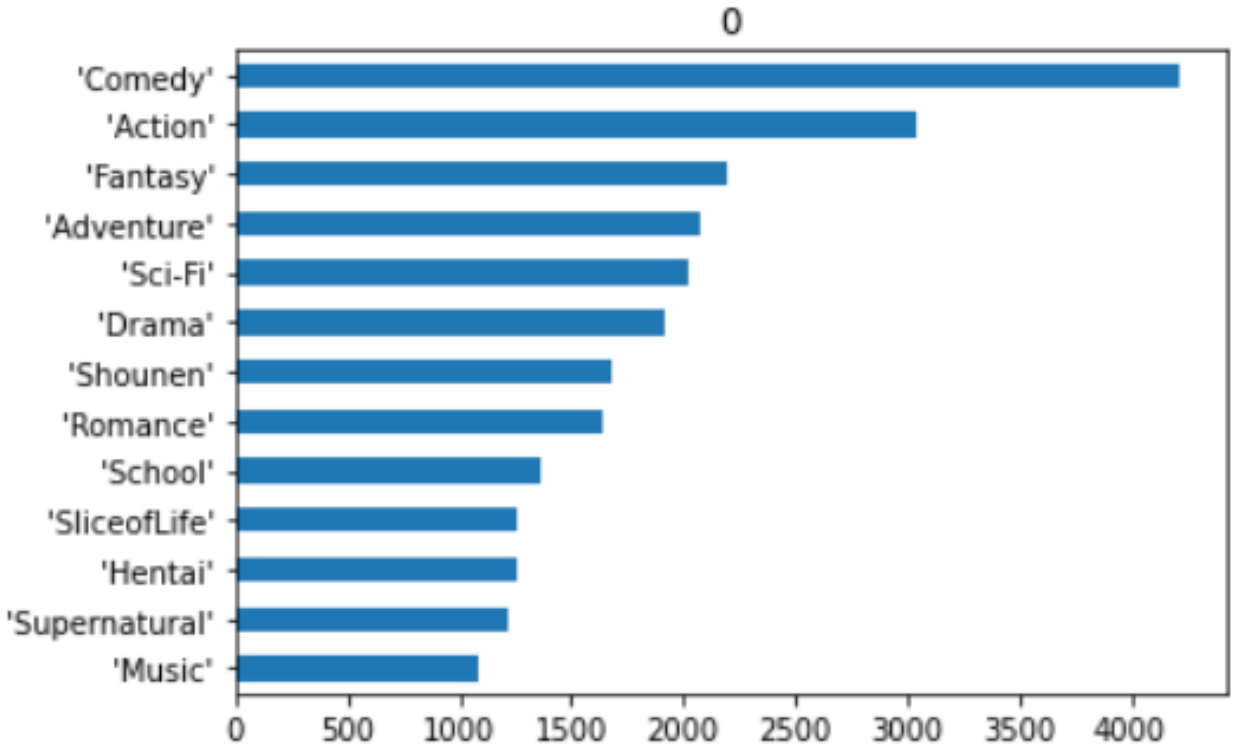
```
temp = df.genres.str.split(",")
print(temp)
import functools
import operator
unique_genres = list(functools.reduce(operator.concat, temp))
num = 0
for item in unique_genres:
  if "[" in item:
    item = item.replace('[','')
  if "]" in item:
    item = item.replace(']','')
  item = item.replace(' ','')
  unique_genres[num] = item
  num = num+1
print(unique_genres)
```

Выделение жанров из строки

Затем с помощью типа данных Counter в огромном списке жанров было подсчитано, сколько раз встречается каждый из жанров. Так как жанров очень много, значения с маленькой встречаемостью были отброшены.

```
from collections import Counter
c = Counter(unique_genres)
c = dict(c)
genres_data = pd.DataFrame.from_dict(c, orient="index")
genres_data
delete_list = []
for i in c:
    if c[i] < 1000:
        delete_list.append(i)
for each in delete_list:
    del c[each]
genres_data2 = pd.DataFrame.from_dict(c, orient="index")
genres_data2
```

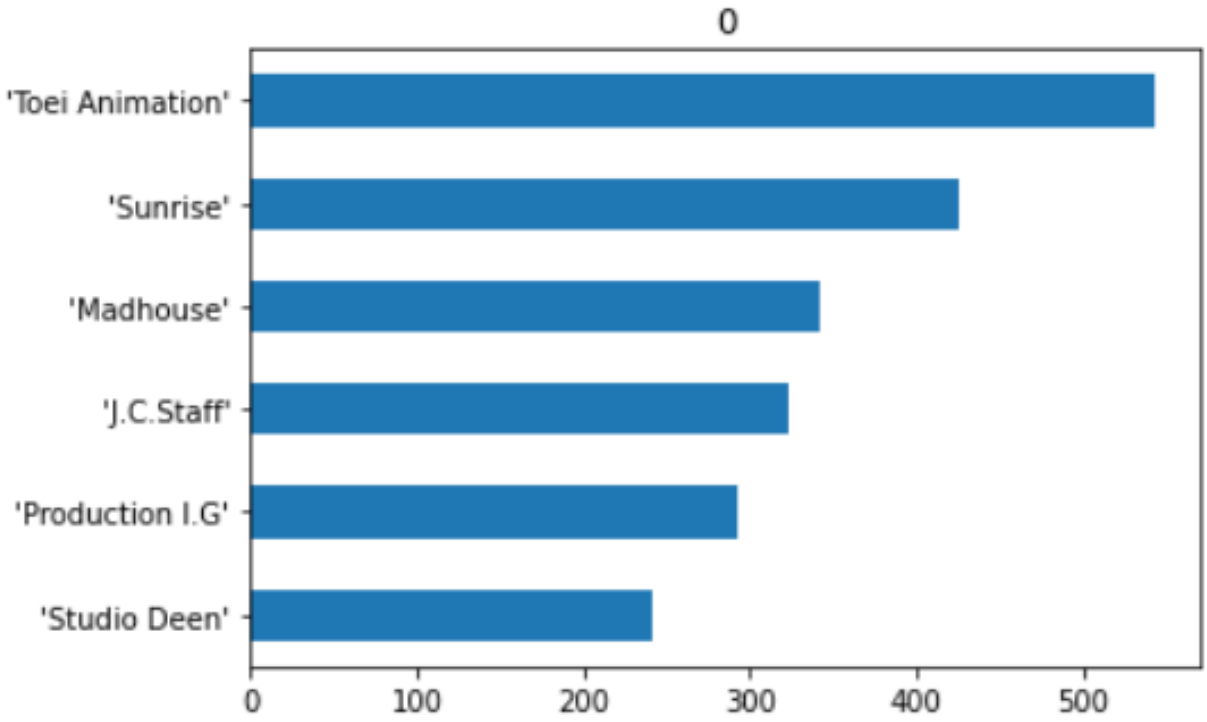
Создание датафрейма с жанрами



Гистограмма распределения жанров

Как видно из диаграммы, несмотря на то, что часть значений была проигнорирована, жанровое разнообразие все равно значительно. Чаще всего встречается комедия и экшн. Это соотносится с тем, что, согласно рейтингу, большинство аниме - сёнены.

Затем изучалось, какие студии чаще всего призводят аниме согласно выбранному датасету. Здесь была аналогичная проблема с тем, что структура, содержащая информацию о студиях, была автоматически преобразована в строку. Решалась она с помощью способа, описанного выше.



Самые продуктивные студии анимеиндустрии

Как видно из гистограммы, было выделено 6 студий, которые производят и выпускают аниме чаще всего. Самая продуктивная студия - Toei Animation.

Сложнее всего было обработать столбец с годом начала выпуска аниме, потому как в этом столбце данные были в разных форматах (строка и вещественный тип данных), а также присутствовали пустые ячейки и прочерки. Необходимо было выделить из этого года.

```
from datetime import datetime
import math
dates = df.aired_from
dates_list = []
num = 0
for i in dates:
  if type(i) == float:
    is_nan = math.isnan(i)
    if is_nan == False:
      i = datetime.fromtimestamp(i).strftime('%Y')
      dates_list.append(i)
  if type(i) == str and i !="-":
    i = i[:4]
    dates_list.append(i)
  num = num+1
```

Выделение года из даты

Так как годов довольно много, для удобства восприятия они были выделены в десятилетия.

```
year_data.plot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f42e2751890>
```

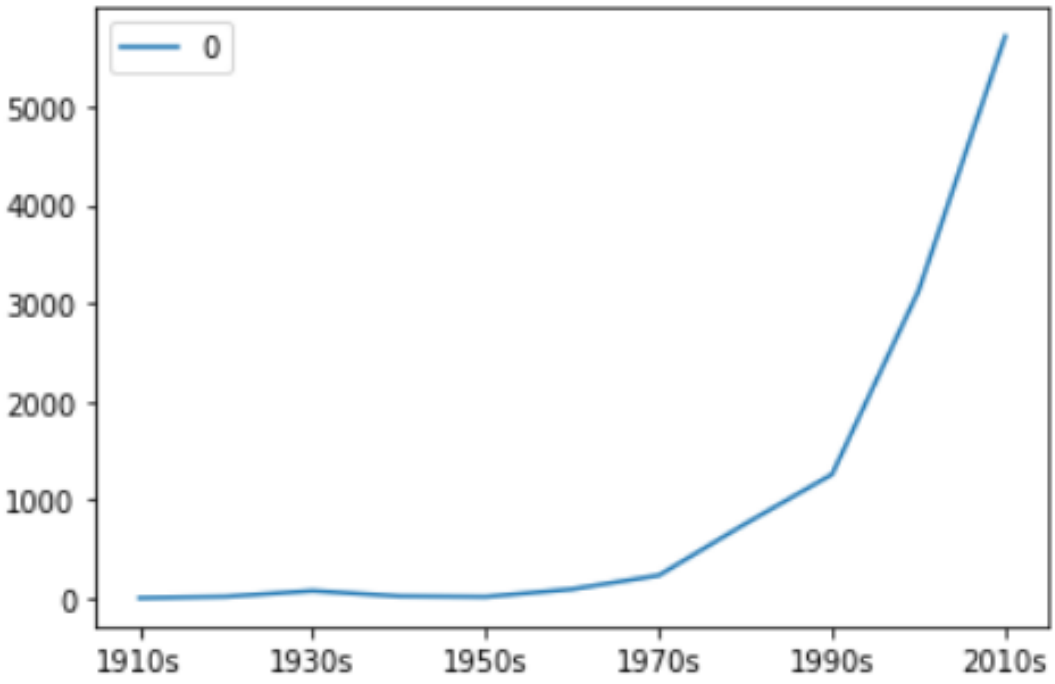


График выпуска аниме

Из графика видно, что с семидесятых годов количество выпущенных аниме стремительно растет. Возможно, это связано с развитием киноиндустрии по всему миру.