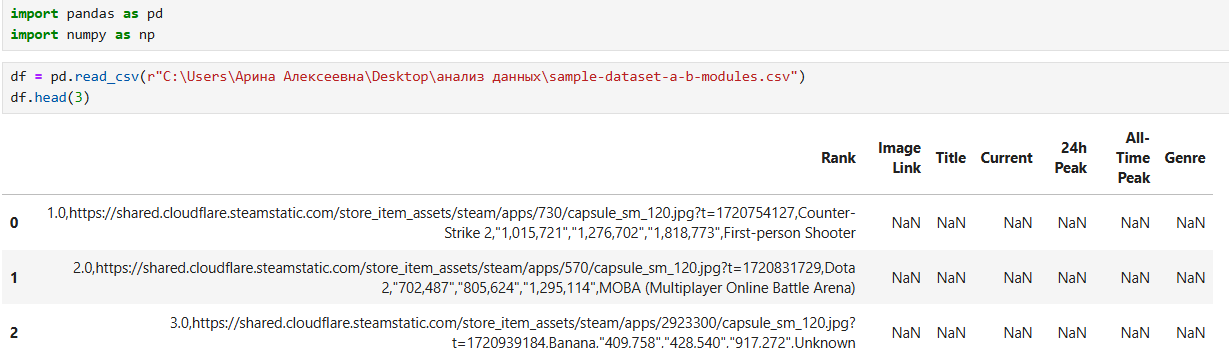
**Отчет**

**Модуль\_А «Предобработка количественных данных»**

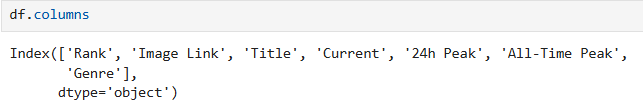
1. **Сбор данных**

Загрузка и импорт необходимых для работы пакетов, загрузка и просмотр данных из csv файла

****

Видно, что данные расположены все в одном столбце Rank с разделителями «,» и «””», где «,» является символом в данных.

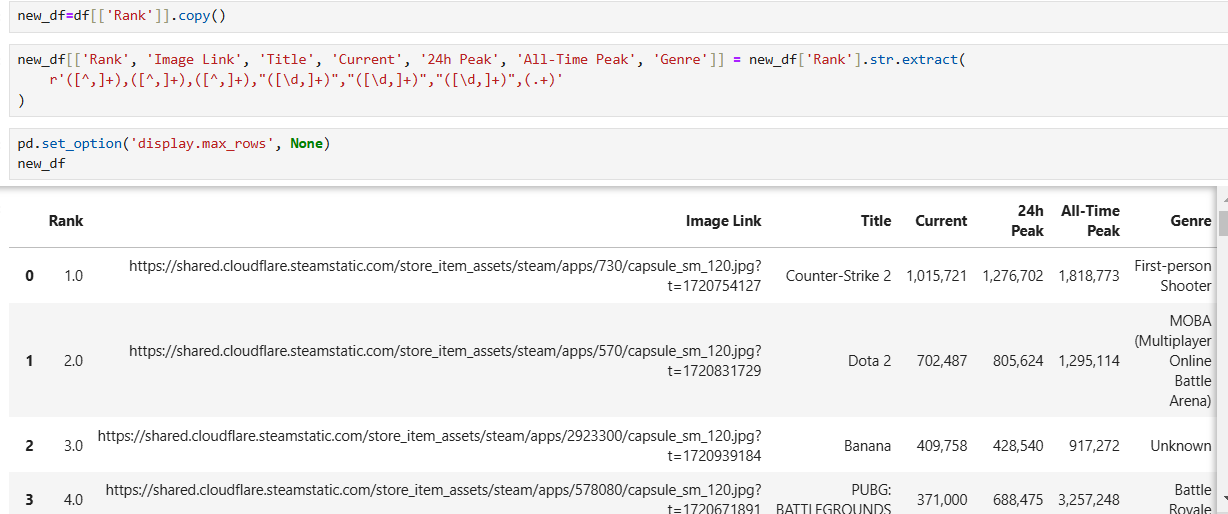
Просматриваю список атрибутов датафрейма



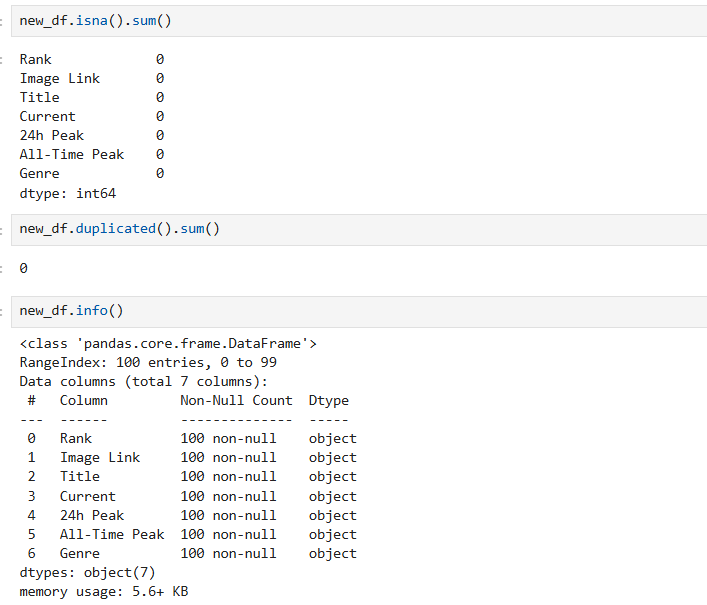
Следует сформировать датафрейм, распределив данные к соответствующим им столбцам.

Используя методы библиотеки pandas, формирую независимый датафрейм,с помощью регулярного выражения к строкам столбца «Rank», извлекаю группы значений.

Применяю опцию для просмотра всех строк, чтобы проверить все ли данные корректно распределились.

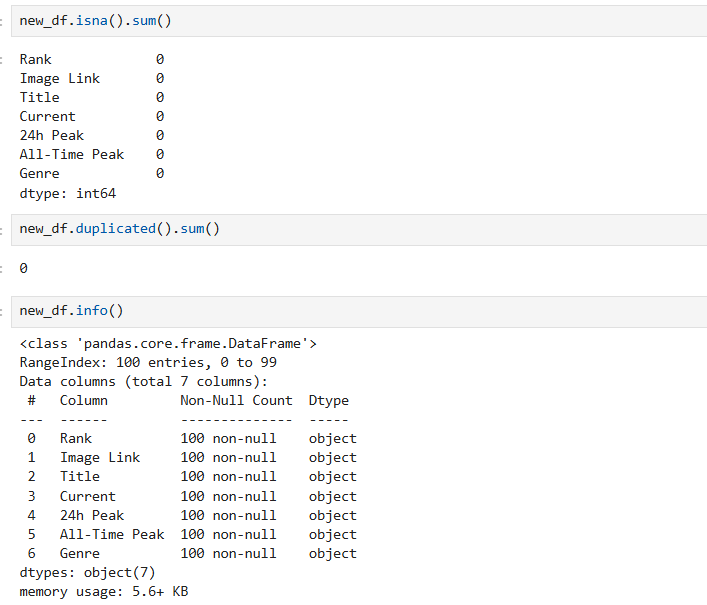


1. **Предобработка данных**

****

Важно исключить ошибки и аномалии, если имеются.

Проверяю данные на наличие пропусков и дубликатов, так как они могут ухудшить качество анализа в дальнейшем. Видно, что пропуски и дубликаты отсутствуют.

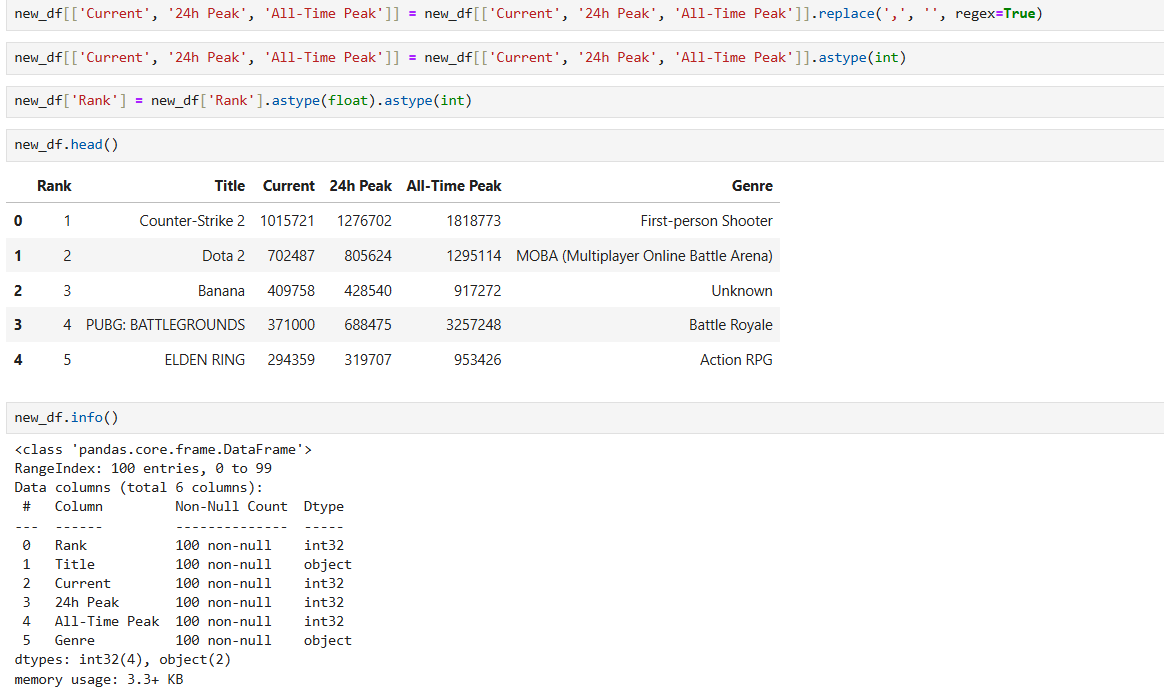
****

Оцениваю общую информацию о данных. Размер датафрейма равен семи столбцам и ста строкам.

Тип данных object, а так как в столбцах есть количественные значения, следует провести преобразование типов данных.

Разные форматы записи чисел могут привести к ошибкам при преобразовании, избавляюсь от разделителей в столбцах под индексами 3-5, а затем преобразую в целочисленный тип int.

Также преобразую нулевой по индексу столбец, в целочисленный тип так как данные- нумерация строк, а также это экономит ресурсы памяти.

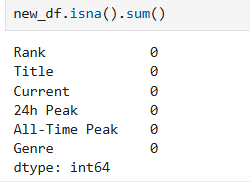


Преобразование прошло успешно, так же сократился объем памяти.

Оценка качества данных, чтобы проверить ничего ли не упущено в предобработке.

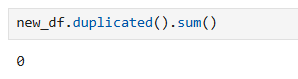
Для этого проведена проверка на:

Полноту(пропуски)



Единообразие (соответствие типов данных)

Достоверность (наличие дубликатов)

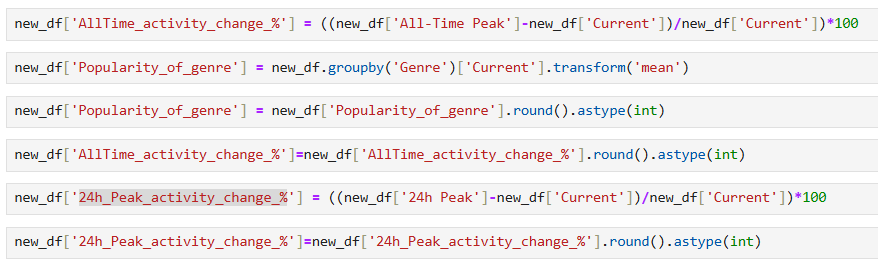


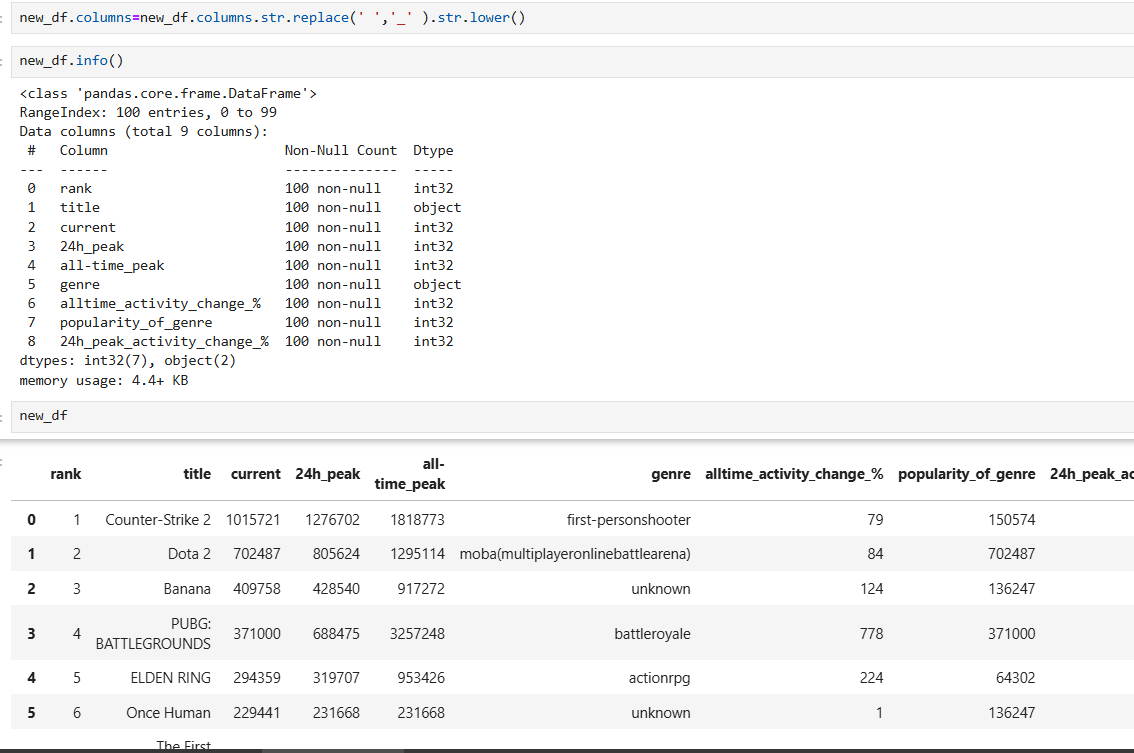
1. **Описание структуры набора данных:**

После проверки качества данных выявляю атрибуты для удаления или добавления.

Избавляюсь от столбца «Image Link», так как он содержит ссылки на изображения, то есть не особо информативен и не поможет при анализе для выявления динамики перемещения человекопотока по играм.

Добавляю новые столбцы – изменение активности в игре за все время относительно текущего, изменение активности в игре за все время относительно суток. Так как эти данные будут полезны при анализе и помогут выдвинуть гипотезы относительно динамики перемещения человекопотока по играм.





Привожу названия столбцов к единому нижнему регистру и заменяю пробел нижним подчеркиванием. Просматриваю общие сведение о датасете после предобработки. Типы данных корректны, пропусков нет, количество столбцов = 9, количество строк = 100.



Экспортирую таблицу после предобработки

*Описание атрибутов*:

rank – номер игры

tittle – название игры

current – текущий показатель онлайн пользователей в игре

24h\_peak – показатель онлайн-пользователей в игре за последние 24 часа

all-time\_peak-показатель онлайн-пользователей в игре за всё время

genre – жанр игры

all-time\_activity\_change\_% - изменение показателя онлайн-активности пользователей в игре за всё время (в процентах)

popularity\_of\_genre – популярность жанра игры

24h\_activity\_change\_% - изменение показателя онлайн-активности пользователей в игре за сутки (в процентах)