Основы программирования в Python
Проект: описание работы
Стихеева Дарья

План

Проект посвящен извлечению и анализу данных с сайта для продажи настольных игр. Содержательная задача - узнать средние показатели по играм в интересующем нас жанре и определить лучшие предложения. Среди показателей, по которым мы будем выявлять подходящие нам товары, выделим следующие.

- Во-первых, это, конечно, цена настолки: предпочтительнее выбрать игру подешевле.
- Во-вторых, это примерное время, которое, по заявлениям производителей, игрокам придется потратить на игру: учитывая обычную высокую цену настолок, нам лучше выбрать игры, требующие много времени.
- Наконец, в-третьих, нас будет интересовать максимальное количество игроков, предусмотренное в настолке. Предположим, что мы хотим играть в компании из 5 человек и больше, поэтому при выборе настолки этот показатель также будет иметь для нас значение.

Данные будут собираться с сайта «hobbygames». Поскольку на нем представлено довольно большое количество игр, а нашей задачей является выбрать себе небольшое число наилучшим образом подходящих вариантов, сразу ограничимся только одним жанром настолок - возьмем приключенческие игры и будем собирать и анализировать данные именно по ним.

Таким образом, программа, которую мы получим на выходе, будет собирать информацию о приключенческих настолках с сайта «hobbygames». В частности, она будет собирать назание игры, ее цену, примерное время игры, максимальное число игроков, а также URL этой игры. Данные по всем играм будут представлены в виде таблицы. Затем программа будет анализировать описательные статистики для цены, времени и количества игроков, после чего в итоговой табличке будут представлены данные только по тем играм, которые будут удовлетворять условиям, которые мы зададим (то есть подходящей нам цене, времени и числу игроков).

Результаты

В наш основной датафрейм по итогу вошло 182 игры (число приводится на настоящий момент и может измениться, если какие-то игры выйдут из наличия или, наоборот, там появятся), для каждого из которых программа сохранила название, цену, время, максимальное число игроков и ссылку.

Далее по данному даатфрейму программа также посчитала средние значения. Средняя цена настолок составляет около 2580 рублей. Средняя продолжительность настолок составляет примерно 86 минут. Максимальное число игроков в среднем составляет 5.

На основе данным по этим столбцам программа также построила три гистограммы, на которых мы видим, что распределение цены скошено влево, большая часть настолок стоит менее 5-6 тысяч; распределение продолжительности настолок также скошено влево, однако среднее значение превышает уже больше наблюдений; такую же скошенность имеет и максимальное число игроков, однако этот показатель имеет значительное количество наблюдений, превышающих среднее.

На основе средних и полученных гистограмм мы строим финальный датафрейм, в который входят только игры, удовлетворяющие трем условиям одновременно. Во-первых, так как распределение цены сильно скошено влево, мы берем только те игры, цена которых ниже 4000 (берем значение выше среднего, поскольку нам необходимы игры для большой компании, а игры с низкой ценой обычно расчитаны на маленькую). Во-вторых, достаточное число наблюдений, превышающих среднюю продолжительность настолок, позволяет выбрать в качестве второго условия все игры дольше 100 минут. В-третьих, поскольку мы ищем игры на компанию из 5 человек или больше, мы также задаем условие о максимуме игроков больше 4. Таким образом, мы получаем финальный датафрейм, в который попадают только те игры, которые удовлетворяют нашим запросам. На данный момент их получилось 12 штук. Это количество оптимально, поскольку у нас есть из чего выбирать,

при этом вариантов не слишком много, чтобы можно было запутаться в описаниях игр и не суметь ничего подобрать.