Python

pandas

Модуль pandas

- pandas — программная библиотека на языке Python для обработки и анализа данных. Работа pandas с данными строится поверх библиотеки NumPy, являющейся инструментом более низкого уровня. Предоставляет специальные структуры данных и операции для манипулирования числовыми таблицами и временны́ми рядами. Название библиотеки происходит от эконометрического термина «панельные данные», используемого для описания многомерных структурированных наборов информации.

Установка

- Для того, чтобы установить модуль numpy, необходимо открыть консоль
- Win + R cmd Enter
- Далее, в консоли необходимо прописать pip install pandas
- Установка завершена

Что дальше?

- Далее открываем снова jupyter notebook (Инструкция в файле Jupyter Notebook.pdf)
- В первой строке прописываем import numpy as np
  - Import – подключение модуля
  - pandas – модуль
  - As pd – используется для сокращенного пользования модулем

Pandas

Данные

Данные

- Чтобы эффективно работать с pandas, необходимо освоить самую главную структуру данных DataFrame. Без понимания что они из себя представляют, невозможно в дальнейшем проводить качественный анализ.

- Объект DataFrame лучше всего представлять себе в виде обычной таблицы и это правильно, ведь DataFrame является табличной структурой данных. В любой таблице всегда присутствуют строки и столбцы.

Создание таблиц данных

- Мы будем использовать готовые данные, которые есть на сайте https://www.kaggle.com/
- Данные, как правило лежат в формате .csv, и что бы добавить их в программу, необходимо воспользоваться следующим методом:

```
df = pd.read_csv('201809-citibike-tripdata.csv')
```

Создание таблиц данных

Out[3]:

	tripduration	starttime	stoptime	start station id	start station name	start station latitude	start station longitude	end station id	end station name	end station latitude	end station longitude	bikeid	usertype
0	1635	2018-09-01 00:00:05.2690	2018-09-01 00:27:20.6340	252.0	MacDougal St & Washington Sq	40.732264	-73.998522	366.0	Clinton Ave & Myrtle Ave	40.693261	-73.968896	25577	Subscriber
1	132	2018-09-01 00:00:11.2810	2018-09-01 00:02:23.4810	314.0	Cadman Plaza West & Montague St	40.693830	-73.990539	3242.0	Schermerhorn St & Court St	40.691029	-73.991834	34377	Subscriber
2	3337	2018-09-01 00:00:20.6490	2018-09-01 00:55:58.5470	3142.0	1 Ave & E 62 St	40.761227	-73.960940	3384.0	Smith St & 3 St	40.678724	-73.995991	30496	Subscriber
3	436	2018-09-01 00:00:21.7460	2018-09-01 00:07:38.5830	308.0	St James Pl & Oliver St	40.713079	-73.998512	3690.0	Park Pl & Church St	40.713342	-74.009355	28866	Subscriber
4	8457	2018-09-01 00:00:27.3150	2018-09-01 02:21:25.3080	345.0	W 13 St & 6 Ave	40.736494	-73.997044	380.0	W 4 St & 7 Ave S	40.734011	-74.002939	20943	Customer
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1877879	369	2018-09-30 00:27:25.9840	2018-09-30 00:33:35.7070	3601.0	Sterling St & Bedford Ave	40.662706	-73.956912	3631.0	Crown St & Bedford Ave	40.666563	-73.956741	32976	Subscriber
1877880	191	2018-09-30 00:30:30.1850	2018-09-30 00:33:42.1090	3601.0	Sterling St & Bedford Ave	40.662706	-73.956912	3631.0	Crown St & Bedford Ave	40.666563	-73.956741	15595	Subscriber
1877881	1442	2018-09-30 08:10:03.1790	2018-09-30 08:34:05.3870	3601.0	Sterling St & Bedford Ave	40.662706	-73.956912	471.0	Grand St & Havemeyer St	40.712868	-73.956981	28646	Subscriber
1877882	453	2018-09-30 12:20:13.6830	2018-09-30 12:27:46.9140	3601.0	Sterling St & Bedford Ave	40.662706	-73.956912	3584.0	Eastern Pkwy & Franklin Ave	40.670777	-73.957680	34272	Subscriber
		2018-09-30	2018-09-30		Sterling St				Clinton St &				

- Как видно на данном скрине – здесь есть очень много данных (строк и столбцов)
- Все эти данные и представляют собой таблицу.

Отдельные столбцы данных

- При анализе данных, нам, как правило необходимо вытаскивать значения отдельных таблиц.
- 1 способ: обращение к самому столбцу данных

```
print(df.tripduration)
```

```
0          1635
1           132
2          3337
3           436
4          8457
           ...
1877879     369
1877880     191
1877881    1442
1877882     453
1877883    1354
Name: tripduration, Length: 1877884, dtype: int64
```

- 2 способ: превращать данные в список:

```
In [6]: st = df['start station id'].tolist()
        st
```

```
Out[6]: [252.0,
         314.0,
         3142.0,
         308.0,
         345.0,
         3142.0,
         3526.0,
         358.0,
         285.0,
         319.0,
         3509.0,
         3077.0,
         500.0,
         3134.0,
         479.0,
         3558.0,
         3107.0,
         350.0,
         3384.0,
         350.0
```

Домашнее задание

- В приложенном датасете фильмы и сериалы от netflix вам нужно найти:
- 1) Чего больше, фильмов или сериалов и вывести результат
- 2) Фильмов какого года больше всего?
- 3) Проверить есть ли в датасете фильмы Казахстана и России