FGA KOMINFO 2022

program DATA SCIENCE

BY @ABELKRISTANTO

Learning Path

FGA KOMINFO 2022

CLASS PROGRAM 3

21/6/22

DATA MANIPULATION

Disini kita akan fokus mempelajari peran pandas sebagai data manipulasi di python

23/6/22

PROJECT CLEANSING

Disini kita akan berfokus untuk dalam penerapan proses cleansing dalam data industri

27/6/22

RECAP

Disini kita akan mengadakan diskusi terkait pembelajaran yang dilakukan sebelumnya

@abelkristanto

FGA KOMINFO 2022

Rule KELAS

Selama kelas diharapkan memenuhi kriteria berikut ini sebelum mulai!

ACTIVE

FEEL FREE TO ASK

CALM AND LEARN

INTERMEZZO

LOCAL

1. Install Anaconda, jika sudah klik Anancoda Prompt dibuka as administrator
2. Ketikan berikut ini dihalaman prompt dengan update pip dengan:

```
python -m pip install --upgrade pip
```

4. Install jupyter notebook:

```
python -m pip install jupyter
```

5. Buka dengan mengetikan:

```
jupyter notebook
```

ONLINE

Jika kamu ingin akses secara online untuk pembelajaran python dapat melalui link dibawah ini ya!

BIT.LY/BELAJARBERSAMAKOHKRIS

welcome to

DATA MANIPULATION

Disini kita akan berfokus untuk library pandas

@abelkristanto

table of CONTENT

Disini kita akan berfokus untuk library pandas

FGA KOMINFO 2022

FGA KOMINFO 2022

Apa itu PANDAS

untuk melakukan analisis dan pengolahan data dari menengah sampai besar

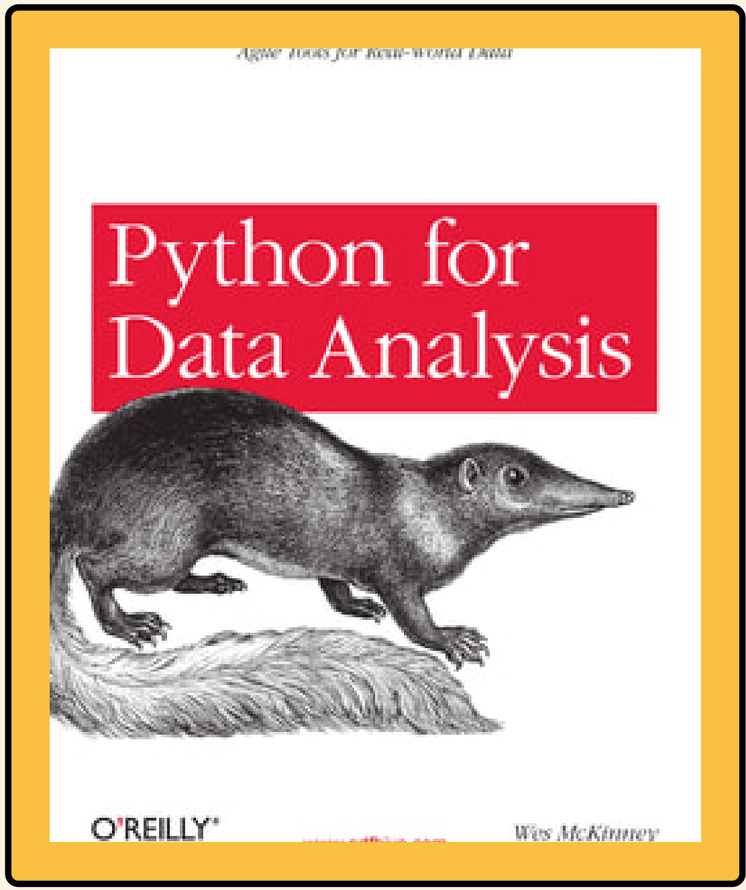
@abelkristanto

INTERMEZZO

POJOK BUKU

Ditulis oleh **Wes McKinney** yang berisikan intisari dalam melakukan data manipulasi dimulai dari perkenalan library terkait (diantaranya numpy, pandas dan environment IPython)

Cocok untuk pemula untuk mempelajari data analisis di python



@abelkristanto

SERIES

Satu kolom bagian dari tabel dataframe yang merupakan 1 dimensional numpy array sebagai basis datanya, terdiri dari 1 tipe data (integer, string, float, dll).

DATAFRAME

Gabungan dari Series, berbentuk rectangular data yang merupakan tabel spreadsheet itu sendiri (karena dibentuk dari banyak Series, tiap Series biasanya punya 1 tipe data, yang artinya 1 dataframe bisa memiliki banyak tipe data).

attribute

PANDAS

1. **.info()** untuk mengecek data types, berapa yang non null, dan sebagainya, cocok untuk DataFrame
2. **.shape** digunakan untuk mengetahui berapa baris dan kolom, hasilnya dalam format tuple (baris, kolom).
3. **.dtypes** digunakan untuk mengetahui tipe data di tiap kolom.
4. **.astype(type_data)** untuk convert tipe data berdasarkan tipe data seperti: float, int, str, numpy.float, numpy.int ataupun numpy.datetime.
5. **.loc** digunakan slice dataframe atau series berdasarkan nama kolom dan/atau nama index.
6. **.iloc** digunakan untuk slice dataframe atau series berdasarkan index kolom dan/atau index.
7. dan **sebagainya**

@abelkristanto

Teknik Baca DATASET

- **.read_csv(....., sep="...")** merupakan cara membaca file yang value-nya dipisahkan oleh comma (default), terkadang pemisah value-nya bisa di set ‘\t’ untuk file tsv (tab separated values).
- **.read_excel(.....)** digunakan untuk membaca file excel menjadi dataframe pandas.
- **.read_json(.....)** digunakan untuk membaca URL API yang formatnya JSON dan mengubahnya menjadi dataframe
- dan **sebagainya**.

Teknik Simpan DATASET

- **.to_csv()** merupakan untuk export dataframe kembali ke csv atau tsv.
- **.to_excel()** merupakan untuk export dataframe menjadi file excel.
- dan **sebagainya**.

INDEXING

Key identifier dari tiap row/column untuk Series atau Dataframe (sifatnya tidak mutable untuk masing-masing value tapi bisa diganti untuk semua value sekaligus). Jika tidak disediakan, pandas akan membuat kolom index default secara otomatis sebagai bilangan bulat (integer) dari 0 sampai range jumlah baris data tersebut.

SLICING

Cara untuk melakukan **filter** ke dalam dataframe/series berdasarkan kriteria tertentu dari nilai kolomnya ataupun kriteria index-nya.

Dimana memiliki dua tipe:
.iloc (proses slicing berdasarkan index berupa nilai integer tertentu)
.loc (lebih fleksibel, dapat berupa value yang ada didalamnya)

TRANSFORMING

Mengubah dataset yang ada menjadi entitas baru, dapat dilakukan dengan: konversi dari satu data type ke data type yang lain, transpose dataframe, atau yang lainnya. Salah satu featurenya:

- **apply()** digunakan untuk menerapkan fungsi di sepanjang sumbu DataFrame atau pada nilai Seri.
- **applymap()** digunakan untuk menerapkan fungsi ke elemen DataFrame.
- **map()** digunakan untuk mengganti setiap nilai dalam Seri dengan nilai lain.

PENGGABUNGAN

- **.append()** dapat digunakan pada dataframe/series yang ditujukan untuk menambah row-nya saja. Jika di SQL memiliki 2 tabel atau lebih maka dapat menggabungkannya secara vertikal dengan Union.
- **.concat()** dapat digunakan pada dataframe yang ditujukan untuk penggabungan baik dalam row-wise (dalam arah) atau column-wise.
- **.merge()** untuk menggabungkan Series/Dataframe yang bentuknya mirip dengan syntax join di SQL, specify left and right tables, join key, dan how to join (left, right, inner, full outer).
- **.join()** digunakan pada dataframe untuk menggabungkan kedua data dengan set index pada kedua tabel tersebut sebagai join key, tanpa index, hal ini tidak akan berhasil.

FGA KOMINFO 2022

PIVOTING PANDAS

- **.pivot()** pada dataframe dapat dilakukan pada dataframe yang memiliki index tunggal ataupun index-nya adalah multi index.
- **.pivot_table()** pada dataframe seperti melakukan pivot pada tabel tapi juga melakukan groupby dan aggregation (aggfunc) pada level rows sehingga dipastikan tidak ada duplicate index di rows (secara default aggfunc = 'mean').
- **.melt()** digunakan untuk mengembalikan kondisi data yang sudah dilakukan pivot menjadi sebelum pivot.

@abelkristanto

FGA KOMINFO 2022

AGGREGASI

Teknik agregasi diperlukan ketika mau melihat dataset dengan view yang berbeda, bisa set data tersebut akan dikelompokkan seperti apa, yang kemudian juga bisa menerapkan beberapa fungsi atau metode statistik ke hasil group dataset itu untuk mengetahui behavior dari data tersebut secara summary/overview.

Konsep pemahaman sebagai berikut:

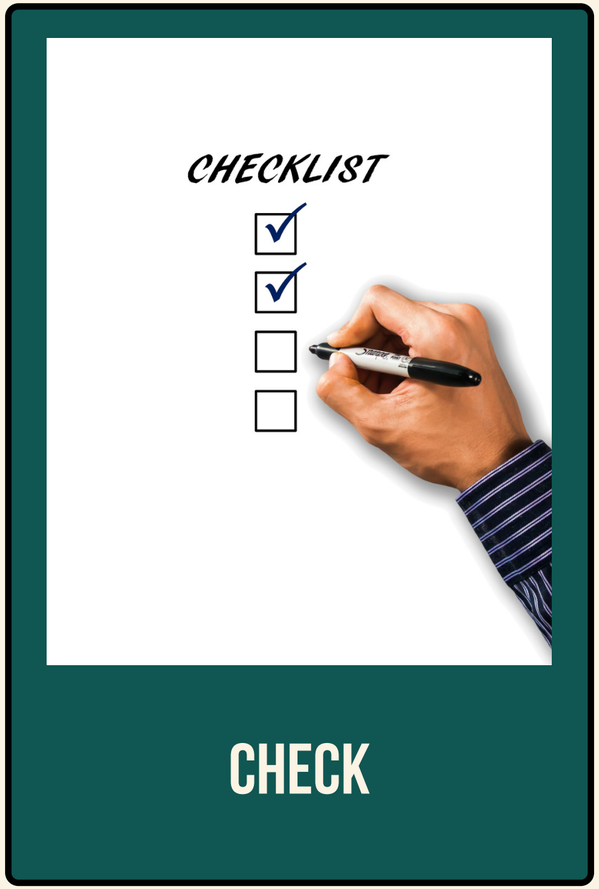
1. Split: melakukan indexing/multi-indexing dengan apa yang di specify as groupby menjadi kelompok
2. Apply: menerapkan fungsi pada masing-masing kelompok tersebut
3. Combine: mengumpulkan semua hasil fungsi dari tiap kelompok kembali menjadi dataframe

@abelkristanto

FGA KOMINFO 2022

Time Series

.to_datetime digunakan untuk men-transform salah satu kolom di dataframe menjadi datetime Pandas dan kemudian set menjadi index.



@abelkristanto

learning BY DOING