DIGITAL TALENT SCHOLARSHIP

THEMATIC ACADEMY

Tema Pelatihan

Pertemuan #06 : Data Understanding (non-visualisasi)

KOMINFO

#JADIJAGOANDIGITAL

Badan Penelitian dan Pengembangan Sumber Daya Manusia

Deskripsi modul dan tujuan pembelajaran

- Modul ini berisi penjelasan mengenai konsep dan teknik pengambilan dan telaah data (*data gathering and understanding*). Teknik-teknik yang dibahas dibatasi pada yang bersifat nonvisual menggunakan statistika. Teknik-teknik visualisasi dijelaskan secara terpisah di modul 07.
- Setelah menyelesaikan modul ini, peserta diharapkan mampu:
  - melakukan pengambilan data untuk proses sains data dari sumber data terbuka, baik secara manual maupun secara programatik menggunakan library Pandas;
  - melakukan telaah data dengan beberapa metode statistika

Referensi

- Materi kuliah Data Mining Fasilkom UI
- Materi kuliah Data Science Fasilkom UI
- Matt Taddy, "Business Data Science", McGraw Hill
- Aggarwal, "Data Mining: The Textbook"
- Joel Grus, "Data Science from Scratch"
- Provost & Fawcett, "Data Science for Business"

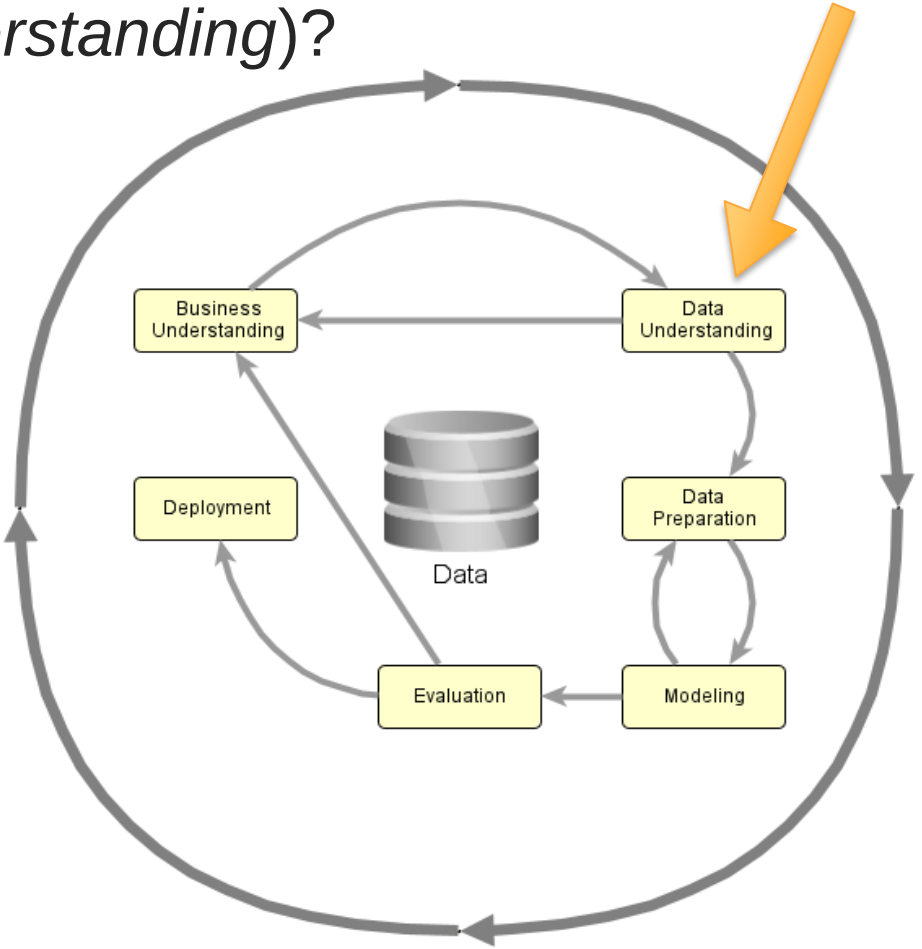
Outline

- Apa itu telaah data (data understanding)?
- Sumber, susunan, tipe, dan model data
- Pengambilan data
- Telaah data dasar

Apa itu Telaah Data (*Data Understanding*)?

Apa itu telaah data (*data understanding*)?

- Dilakukan setelah problem bisnis terdefinisikan sebagai hasil tahapan business understanding.
- Tujuan: mendapatkan gambaran utuh atas data.
- Dilanjutkan ke persiapan data (data preparation), jika pemahaman awal data cukup atau kembali ke business understanding jika definisi permasalahan bisnis harus direvisi.



Mengapa perlu data understanding?

- Data = bahan mentah solusi AI
- Data dari masing-masing sumber belum tentu dapat langsung dipakai karena:
  - maksud dan tujuan data berbeda-beda
  - keadaan asal terpisah-pisah atau justru terintegrasi secara ketat.
  - tingkat kekayaan (*richness*) berbeda-beda
  - tingkat keandalan (*reliability*) berbeda-beda
- Data understanding memberikan gambaran awal tentang:
  - kekuatan data
  - kekurangan dan batasan penggunaan data
  - tingkat kesesuaian data dengan masalah bisnis yang akan dipecahkan
  - ketersediaan data (terbuka/tertutup, biaya akses, dsb.)

Bagian-bagian proses telaah data

- Identifikasi "titik sentuh" data dengan proses bisnis
- Penentuan sumber utama data dan cara aksesnya
- Asesmen nilai tambah bisnis dari data
- Identifikasi sumber data tambahan untuk perbaikan

Sumber, Susunan, Tipe dan Model Data

Sumber data

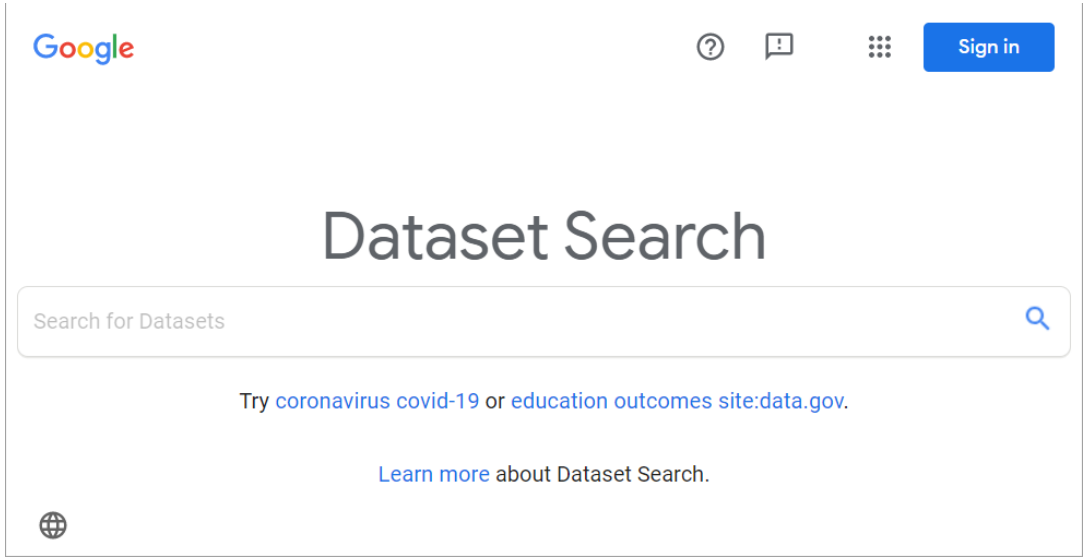
Internal sources	Spreadsheets (Excel, CSV, JSON, etc.)
	Databases: can be queried via SQL, etc.
	Text documents
	Multimedia documents (audio, video)
External sources	Open data repositories
	Public domain web pages

Sumber data daring

- Portal Satu Data Indonesia (https://data.go.id)
- Portal Data Jakarta (https://data.jakarta.go.id)
- Portal Data Bandung (http://data.bandung.go.id)
- Badan Pusat Statistik (https://www.bps.go.id)
- Badan Informasi Geospasial (https://tanahair.indonesia.go.id/)
- UCI Machine Learning repository (https://archive.ics.uci.edu/ml/index.php)
- Kaggle (https://www.kaggle.com/datasets)
- World Bank Open Data (https://data.worldbank.org)
- UNICEF Data (https://data.unicef.org)
- WHO Open Data (https://www.who.int/data)
- IBM Data Asset eXchange (https://developer.ibm.com/exchanges/data/)
- DBPedia (https://www.dbpedia.org/resources/)
- Wikidata (https://www.wikidata.org/) .

Sumber data daring

- Cari via Google Dataset Search: https://datasetsearch.research.google.com



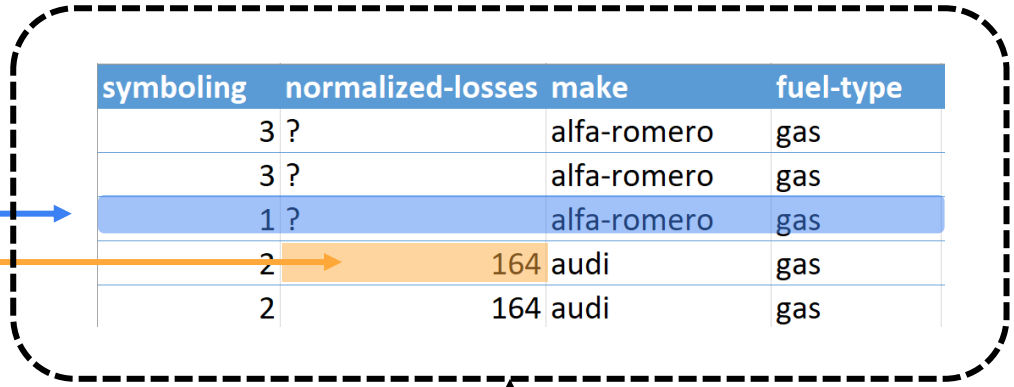
Susunan data

Butir data (*datum*): satuan terkecil data; satu nilai untuk satu variable tertentu

Data: kumpulan butir data yang membawa satu kesatuan makna (mendeskripsikan satu objek) tertentu.

Himpunan data (*dataset*): kumpulan data.

Metadata: data yang menjelaskan data yang lain.



symboling	normalized-losses	make	fuel-type
3	?	alfa-romero	gas
3	?	alfa-romero	gas
1	?	alfa-romero	gas
2	164	audi	gas
2	164	audi	gas

"make":
- tipe: string,
- deskripsi: nama pabrikan merek kendaraan

Tipe data berdasarkan susunannya

	Data terstruktur (structured data)	Data takterstruktur (unstructured data)
Sifat	• Model data terdefinisikan sebelumnya • Format butir data (biasanya) teks. • Antar butir data terbedakan dengan jelas. • Ekstraksi/kueri langsung cukup mudah.	• Model data tidak terdefinisikan sebelumnya • Format butir data (biasanya) teks, citra, suara, video, dan format lainnya. • Antar butir data tidak cukup jelas terbedakan karena ketidakteraturan dan ambiguitas. • Ekstraksi/kueri langsung cukup sulit.
Contoh	Data tabular, data berorientasi objek, *time series*	Data teks dalam dokumen teks bebas, data audio, data video.

Data semi-terstruktur (*semi-structured data*): Data terstruktur yang tidak mengikuti model struktur tabular yang seperti pada basis data relasional, namun tetap mengandung *tags* atau penanda lainnya yang dapat memisahkan elemen-elemen semantik pada data serta mengatur hierarki antara butir-butir datanya.

Tipe butir data (1)

	Nominal/kategorikal	Ordinal	Interval	Rasio
Sifat himpunan asal	Diskret, tidak terurut	Diskret, terurut	Kontinu/numerik, terurut, perbedaan menunjukkan selisih	Kontinu/numerik, terurut, nilai menunjukkan rasio terhadap kuantitas satuan/unit di jenis yang sama
Contoh	Warna (merah, hijau, biru)	Nilai huruf mahasiswa (A, B, C, D, E)	Suhu dalam Celcius, tanggal dalam kalender tertentu	Panjang jalan, suhu dalam Kelvin
Ukuran data menyatakan …	Membership	Membership, comparison	Membership, comparison, difference	Membership, comparison, difference, magnitude
Operasi matematika	=, ≠	=, ≠, <, >	=, ≠, <, >, +, -	=, ≠, <, >, +, -, ×, ÷

Tipe butir data (2)

	Nominal/kategorikal	Ordinal	Interval	Rasio
Representasi nilai tipikal	Modus	Modus, median	Modus, median, rerata aritmetis	Modus, median, rerata aritmetis, rerata geometris, rerata harmonis
Representasi sebaran	Grouping	Grouping, rentang (*range*), rentang antarkuartil	Grouping, rentang (*range*), rentang antarkuartil, varians, simpangan baku	Grouping, rentang (*range*), rentang antarkuartil, varians, simpangan baku, koefisien variasi
Memiliki nol sejati yang menyatakan nilai mutlak terbawah.	Tidak	Tidak	Tidak	Ya

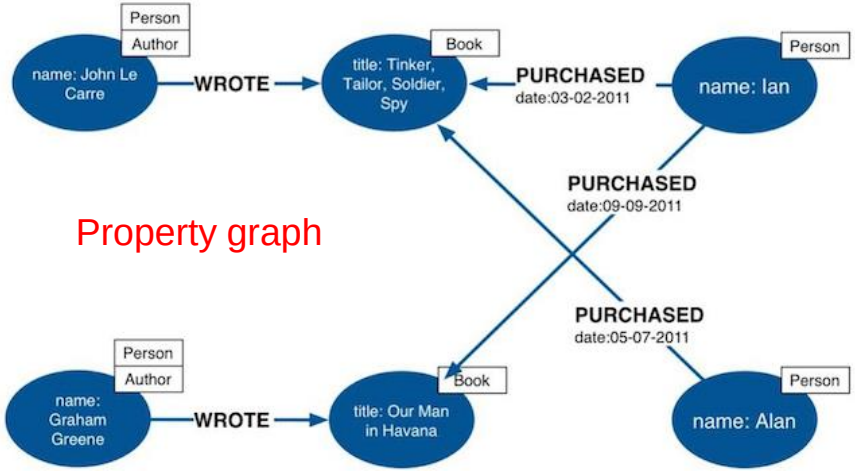
Contoh model data: Tabular

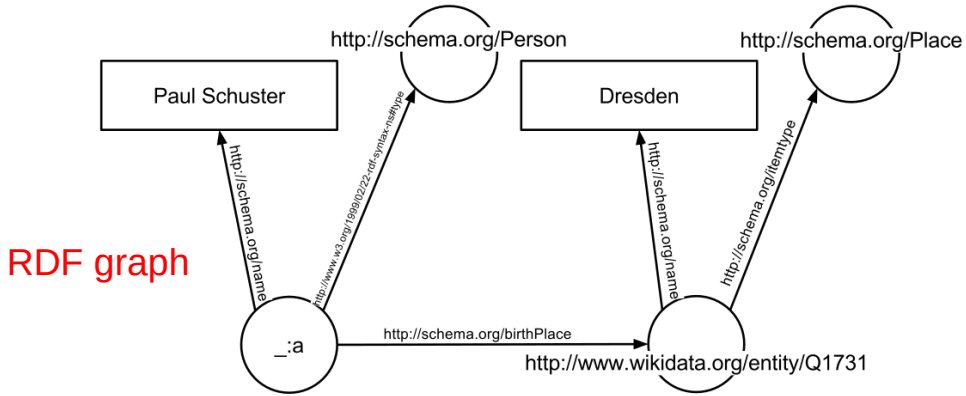
- Terdiri dari N buah rekord (*record*)
- Masing-masing rekord mengandung D buah atribut
- Rekord = baris, *data point*, instans, *example*, transaksi, tupel, entitas, objek, vector fitur.
- Atribut = kolom, *field*, dimensi, fitur.
- Atribut yang sama untuk setiap rekord biasanya diasumsikan memiliki tipe butir data yang sama.
- Struktur dapat bersifat ketat/strict (contoh: basis data relasional) atau longgar/loose (contoh: Excel *spreadsheet*).
- Tergantung keketatan strukturnya, bisa ada bahasa kueri formal untuk mengakses butir-butir data di dalamnya (contoh: SQL).

symboling	normalized-losses	make
3	?	alfa-romero
3	?	alfa-romero
1	?	alfa-romero
2	164	audi
2	164	audi

Contoh model data: Graf/Jejaring

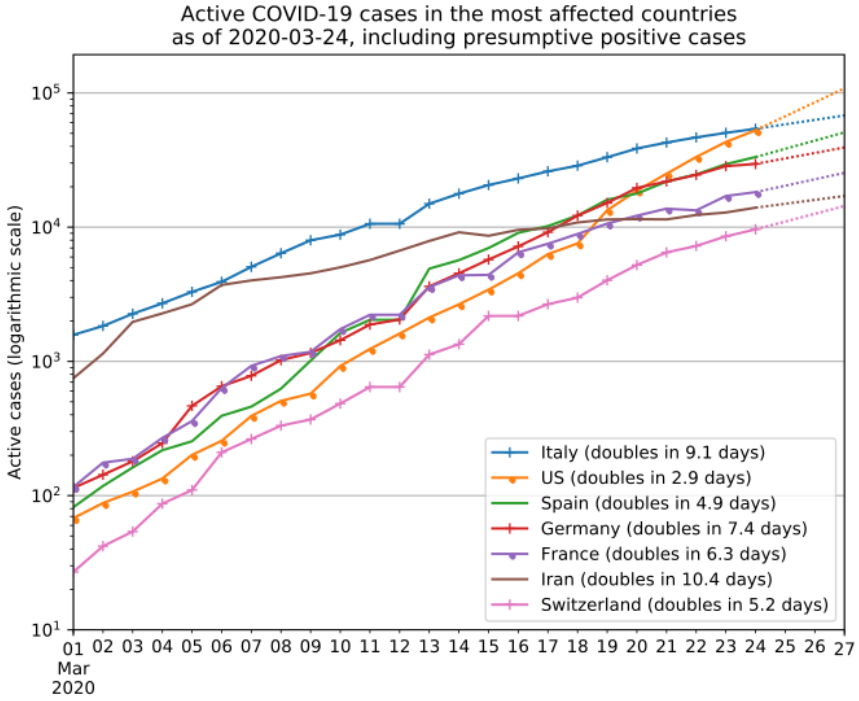
- Tersusun dari simpul-simpul (*nodes*) dan sisi/koneksi antar simpul (*edges*)
- Satu node (biasanya) mewakili satu rekord
- Dapat mengekspresikan relasi antar rekord secara eksplisit.
- Termasuk model data graf adalah model data hierarkis/pohon, model data berorientasi objek (*object-oriented data model*).
- Model data graf modern:
  - *Property graph*
  - *Resource description framework (RDF)*





Contoh model data: Sekuens

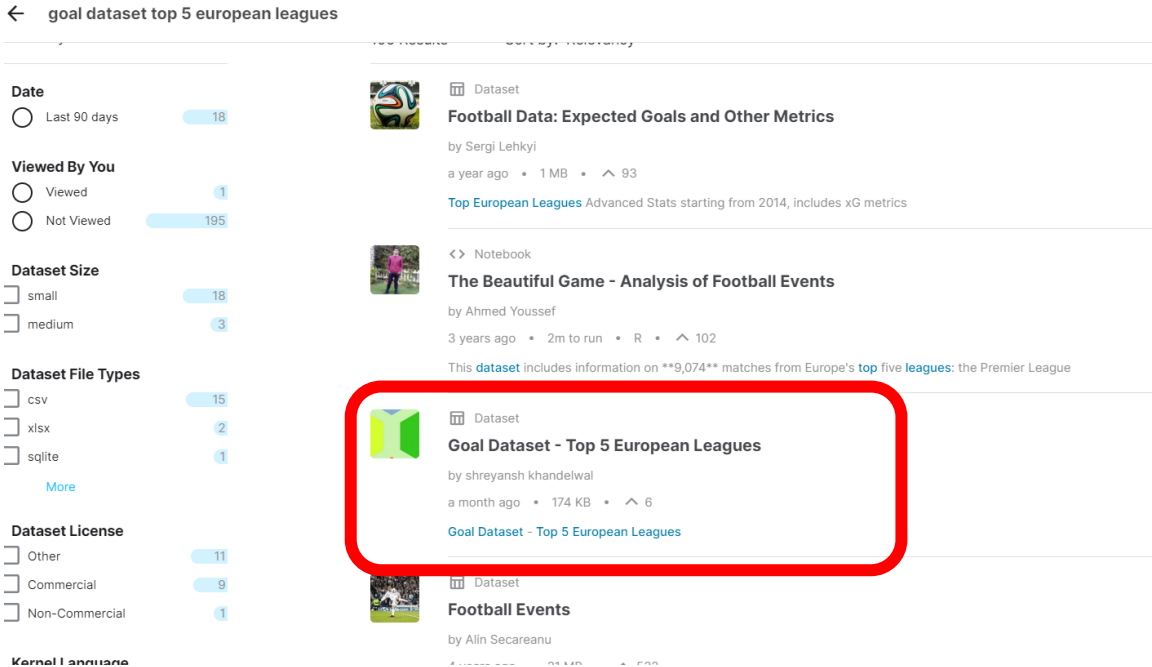
- Tersusun dari rekord-rekord yang terhubung secara sekuensial.
- Contoh: data dari sensor suhu selama suatu rentang waktu.
- Struktur tersirat dari urutan kemunculan rekord
- Rekaman audio dan video dapat dipandang sebagai data sekuens, namun setiap rekordnya sendiri bersifat tidak terstruktur.
- Atribut kontekstual mendefinisikan basis dependensi tersirat. (Contoh: time stamp pada sensor suhu)
- Atribut behavioral: butir-butir data yang nilainya diperoleh dalam suatu konteks tertentu (Contoh: besarnya suhu).
- Jika atribut kontekstualnya adalah waktu/time stamp, maka data sekuens disebut *time series*.

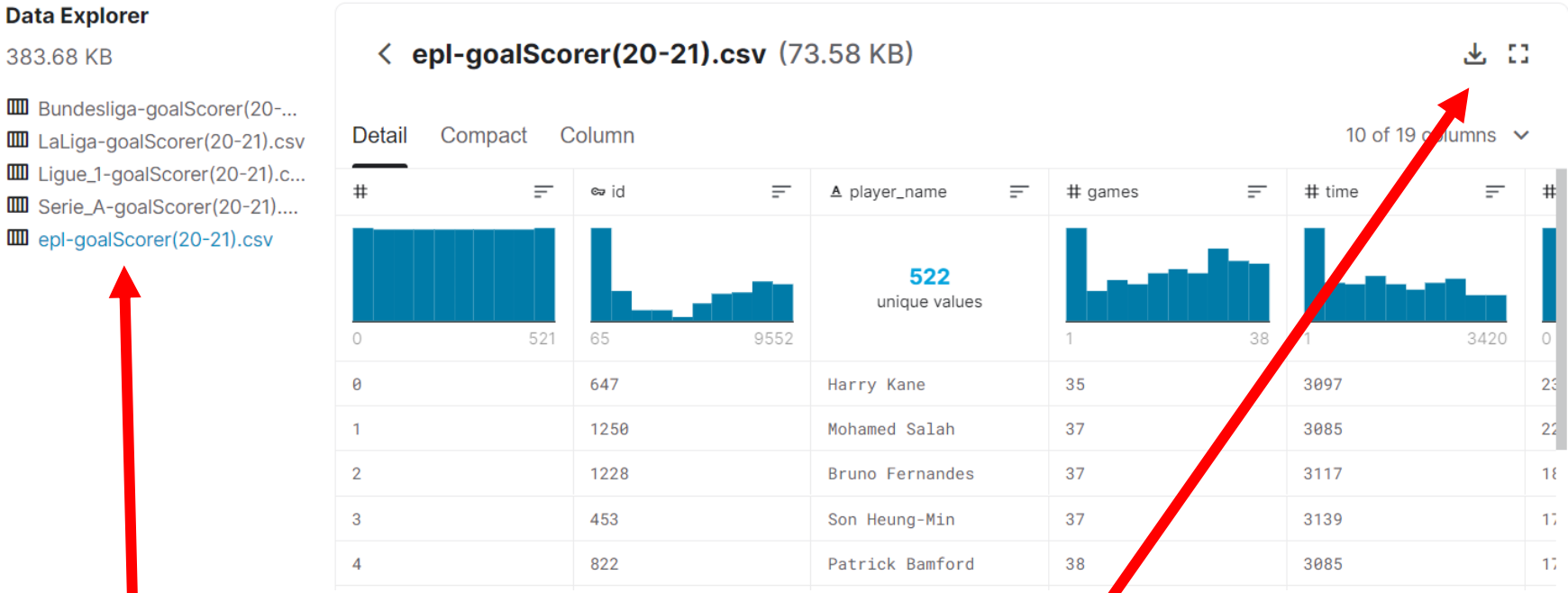


Pengambilan Data

Mengambil data dari Kaggle

- Kita akan mengakses data dari "Goal Dataset – Top 5 European Leagues" dari Kaggle.
- Kunjungi Kaggle.com dan login (buat akun jika perlu)
- Search "goal dataset top 5 European leagues"
- Klik "Goal Dataset – Top 5 European Leagues"





- Di halaman data explorer, pilih "epl-goalScorer (20-21).csv"
- Unduh data dengan mengklik tombol unduh di bagian kanan dan simpan di folder kerja Anda.

Pengambilan data dengan cara lain

- Kaggle dan beberapa layanan data lainnya menyediakan akses melalui API.
- Langkah-langkah mengakses API biasanya melalui proses pembuatan API token/API key yang dirinci di dokumentasi masing-masing layanan.
- Selain API, teknik pengambilan data yang bersifat lanjut mencakup *web scraping* serta akses data langsung dari basis data relasional.

Memuat data ke Pandas (1)

- Nyalakan Jupyter Notebook di folder kerja Anda.
- Buka atau buat baru satu skrip ipynb (Python 3)
- Import pandas dan numpy. (Pastikan sudah terinstal sebelumnya).
- Load file CSV yang diunduh ke dalam sebuah DataFrame
  - Gunakan perintah `read_csv(…)`

```
In [1]: import pandas as pd
        import numpy as np

In [2]: path = "epl-goalScorer(20-21).csv"
        df = pd.read_csv(path)
```

Memuat data ke Pandas (2)

- Method head() dan tail() pada DataFrame membantu kita menampilkan beberapa baris pertama/terakhir dari data yang kita muat.

```
In [4]: df.head(3)
```

Out[4]:

	Unnamed: 0	id	player_name	games	time	goals	xG	assists
0	0	647	Harry Kane	35	3097	23	22.174859	14
1	1	1250	Mohamed Salah	37	3085	22	20.250847	5
2	2	1228	Bruno Fernandes	37	3117	18	16.019454	12

```
In [3]: df.head()
```

Out[3]:

	Unnamed: 0	id	player_name	games	time	goals	xG	assists
0	0	647	Harry Kane	35	3097	23	22.174859	14
1	1	1250	Mohamed Salah	37	3085	22	20.250847	5
2	2	1228	Bruno Fernandes	37	3117	18	16.019454	12
3	3	453	Son Heung-Min	37	3139	17	11.023287	10
4	4	822	Patrick Bamford	38	3085	17	18.401863	7

Telaah Data

Mengungkap tipe-tipe data dari setiap kolom

- Atribut `dtypes` pada DataFrame memberikan tipe data dari setiap kolom.
- Lihat Pandas User Guide untuk detil dari setiap tipe.
- `dtype:object` di akhir output dtypes mewakili `Series` yang merupakan objek Python yang dikembalikan oleh dtypes itu sendiri (bukan bagian dari tipe kolom manapun).

```
In [5]: print(df.dtypes)
```

```
Unnamed: 0          int64
id                  int64
player_name        object
games               int64
time                int64
goals               int64
xG                float64
assists             int64
xA                float64
shots               int64
key_passes          int64
yellow_cards        int64
red_cards           int64
position           object
team_title         object
npg                 int64
npxG              float64
xGChain           float64
xGBuildup         float64
dtype: object
```

Deskripsi statistik data

DataFrame method describe() menampilkan statistik dasar setiap kolom data yang bertipe numerik, mencakup banyaknya data (**count**), rerata aritmetik (**mean**), simpangan baku (**std**), nilai terkecil (**min**), kuartil pertama (**25%**), kuartil kedua/median (**50%**), kuartil ketiga (**75%**), dan nilai terbesar (**max**).

```
In [6]: df.describe()
```

Out[6]:

	Unnamed: 0	id	games	time	goals	xG	assists	xA	shots	key_passes	yellow_cards	red_cards
count	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000
mean	260.500000	4380.932950	19.643678	1420.068966	1.862069	2.000806	1.289272	1.376029	17.379310	12.963602	2.061303	0.091954
std	150.832689	3281.776121	11.619836	1031.604819	3.338851	3.317946	2.083350	1.886510	21.572664	16.164361	2.203661	0.295800
min	0.000000	65.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	130.250000	839.750000	10.000000	470.250000	0.000000	0.074668	0.000000	0.049245	2.000000	1.000000	0.000000	0.000000
50%	260.500000	4627.000000	21.000000	1342.000000	1.000000	0.737295	0.000000	0.691122	10.000000	7.000000	2.000000	0.000000
75%	390.750000	7690.500000	30.000000	2319.000000	2.000000	2.053378	2.000000	2.050509	23.750000	19.000000	3.000000	0.000000
max	521.000000	9552.000000	38.000000	3420.000000	23.000000	22.174859	14.000000	11.474996	138.000000	95.000000	12.000000	2.000000

Deskripsi statistik data

Gunakan `describe(include='all')` jika ingin menampilkan juga statistik kolom yang bertipe non-numerik, mencakup juga berapa banyak nilai unik dalam kolom (**unique**), nilai modus (**top**), serta frekuensi modus (**freq**).

```
In [7]: df.describe(include='all')
```

Out[7]:

	Unnamed: 0	id	player_name	games	time	goals	xG	assists	xA	shots	key_passes	yellow_card	red_cards	position	team_title	npg	npxG	xG
count	522.000000	522.000000	522	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.00000	522.000000	522	522	522.000000	522.000000	522.0
unique	NaN	NaN	522	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Nal	NaN	14	28	NaN	NaN	
top	NaN	NaN	Alex McCarthy	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Nal	NaN	M S	West Bromwich Albion	NaN	NaN	
freq	NaN	NaN	1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	Nal	NaN	106	28	NaN	NaN	
mean	260.500000	4380.932950	NaN	19.643678	1420.068966	1.862069	2.000806	1.289272	1.376029	17.379310	12.963602	2.06130	0.091954	NaN	NaN	1.668582	1.821450	5.66
std	150.832689	3281.776121	NaN	11.619836	1031.604819	3.338851	3.317946	2.083350	1.886510	21.572664	16.164361	2.20366	0.295800	NaN	NaN	2.909929	2.931176	5.60
min	0.000000	65.000000	NaN	1.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000	0.000000	NaN	NaN	0.000000	0.000000	0.00
25%	130.250000	839.750000	NaN	10.000000	470.250000	0.000000	0.074668	0.000000	0.049245	2.000000	1.000000	0.00000	0.000000	NaN	NaN	0.000000	0.074668	1.19
50%	260.500000	4627.000000	NaN	21.000000	1342.000000	1.000000	0.737295	0.000000	0.691122	10.000000	7.000000	2.00000	0.000000	NaN	NaN	0.500000	0.715585	4.25
75%	390.750000	7690.500000	NaN	30.000000	2319.000000	2.000000	2.053378	2.000000	2.050509	23.750000	19.000000	3.00000	0.000000	NaN	NaN	2.000000	1.945799	8.30
max	521.000000	9552.000000	NaN	38.000000	3420.000000	23.000000	22.174859	14.000000	11.474996	138.000000	95.000000	12.00000	2.000000	NaN	NaN	19.000000	19.130183	28.96

Fungsi statistik dalam Pandas

`count`	Number of non-NA observations
`sum`	Sum of values
`mean`	Mean of values
`mad`	Mean absolute deviation
`median`	Arithmetic median of values
`min`	Minimum
`max`	Maximum
`mode`	Mode
`abs`	Absolute Value
`prod`	Product of values

`std`	Bessel-corrected sample standard deviation
`var`	Unbiased variance
`sem`	Standard error of the mean
`skew`	Sample skewness (3rd moment)
`kurt`	Sample kurtosis (4th moment)
`quantile`	Sample quantile (value at %)
`cumsum`	Cumulative sum
`cumprod`	Cumulative product
`cummax`	Cumulative maximum
`cummin`	Cumulative minimum

Contoh fungsi statistik setiap kolom (yang *applicable*)

```
In [8]: df.mean()
```

```
Out[8]: Unnamed: 0       260.500000
        id              4380.932950
        games             19.643678
        time            1420.068966
        goals              1.862069
        xG                 2.000806
        assists            1.289272
        xA                 1.376029
        shots             17.379310
        key_passes        12.963602
        yellow_cards       2.061303
        red_cards          0.091954
        npg                1.668582
        npxG               1.821450
        xGChain            5.663368
        xGBuildup          3.455060
        dtype: float64
```

```
In [9]: df.sum()
```

```
Out[9]: Unnamed: 0                                                135981
        id                                                       2286847
        player_name     Harry KaneMohamed SalahBruno FernandesSon Heun...
        games                                                      10254
        time                                                      741276
        goals                                                        972
        xG                                                    1044.420572
        assists                                                      673
        xA                                                    718.287269
        shots                                                        9072
        key_passes                                                   6767
        yellow_cards                                                 1076
        red_cards                                                      48
        position        FF M SM SF M SF SF SF SFM SF M SF SF SF SF SF ...
        team_title      TottenhamLiverpoolManchester UnitedTottenhamLe...
        npg                                                          871
        npxG                                                    950.7971
        xGChain                                               2956.278233
        xGBuildup                                             1803.541131
        dtype: object
```

Contoh fungsi statistik setiap kolom (yang *applicable*)

```
In [10]: df.median()
```

```
Out[10]: Unnamed: 0      260.500000
         id             4627.000000
         games            21.000000
         time           1342.000000
         goals             1.000000
         xG                0.737295
         assists           0.000000
         xA                0.691122
         shots            10.000000
         key_passes        7.000000
         yellow_cards      2.000000
         red_cards         0.000000
         npg               0.500000
         npxG              0.715585
         xGChain           4.252738
         xGBuildup         2.656397
         dtype: float64
```

```
In [12]: df.std()
```

```
Out[12]: Unnamed: 0      150.832689
         id             3281.776121
         games            11.619836
         time           1031.604819
         goals             3.338851
         xG                3.317946
         assists           2.083350
         xA                1.886510
         shots            21.572664
         key_passes       16.164361
         yellow_cards      2.203661
         red_cards         0.295800
         npg               2.909929
         npxG              2.931176
         xGChain           5.600249
         xGBuildup         3.376584
         dtype: float64
```

```
In [14]: df.quantile(0.75)
```

```
Out[14]: Unnamed: 0      390.750000
         id             7690.500000
         games            30.000000
         time           2319.000000
         goals             2.000000
         xG                2.053378
         assists           2.000000
         xA                2.050509
         shots            23.750000
         key_passes       19.000000
         yellow_cards      3.000000
         red_cards         0.000000
         npg               2.000000
         npxG              1.945799
         xGChain           8.308002
         xGBuildup         5.254647
         Name: 0.75, dtype: float64
```

Value_counts

- value_counts() menghasilkan frekuensi setiap nilai unik di dalam kolom.
- Yang tertinggi count-nya adalah merupakan modus pada kolom tersebut.

```
In [18]: df['team_title'].value_counts()

Out[18]: West Bromwich Albion              28
         Everton                           28
         Fulham                            27
         Wolverhampton Wanderers           27
         Southampton                       27
         Sheffield United                  27
         Manchester United                 27
         Liverpool                         27
         Leicester                         27
         Brighton                          26
         Arsenal                           26
         Newcastle United                  26
         Chelsea                           25
         Burnley                           25
         Tottenham                         24
         Manchester City                   24
         Crystal Palace                    24
         West Ham                          23
         Leeds                             23
         Aston Villa                       23
         West Bromwich Albion,West Ham      1
         Everton,Southampton                1
         Arsenal,West Bromwich Albion       1
         Chelsea,Fulham                     1
         Aston Villa,Chelsea                1
         Arsenal,Newcastle United           1
         Liverpool,Southampton              1
         Arsenal,Brighton                   1
         Name: team_title, dtype: int64
```

Korelasi Pearson antara kolom-kolom numerik

- Method corr() menghasilkan tabel korelasi Pearson antar kolom-kolom numerik.
- Rentang nilai: antara -1 dan 1.
- -1 = korelasi negatif, 0 = tidak ada korelasi linear, +1 = korelasi positif.

```
In [23]: df.loc[:,'games':].corr()
```

Out[23]:

	games	time	goals	xG	assists	xA	shots	key_passes	yellow_cards	red_cards	npg	npxG	xGChain	xGBu
games	1.000000	0.944591	0.439730	0.463869	0.504168	0.562806	0.599164	0.617867	0.565963	0.160326	0.437110	0.465546	0.726598	0.6
time	0.944591	1.000000	0.398930	0.411203	0.473555	0.516638	0.529534	0.575065	0.592223	0.186333	0.392631	0.408231	0.703801	0.7
goals	0.439730	0.398930	1.000000	0.932798	0.617490	0.607330	0.873363	0.567752	0.097151	0.053679	0.971591	0.905710	0.727953	0.2
xG	0.463869	0.411203	0.932798	1.000000	0.636205	0.627495	0.910214	0.570488	0.093761	0.048815	0.894286	0.979218	0.763909	0.2
assists	0.504168	0.473555	0.617490	0.636205	1.000000	0.885850	0.721220	0.835299	0.209349	-0.021444	0.587316	0.615503	0.752587	0.4
xA	0.562806	0.516638	0.607330	0.627495	0.885850	1.000000	0.759568	0.946506	0.243912	0.006284	0.585152	0.611100	0.814487	0.5
shots	0.599164	0.529534	0.873363	0.910214	0.721220	0.759568	1.000000	0.743370	0.249957	0.073932	0.852989	0.901386	0.843152	0.4
key_passes	0.617867	0.575065	0.567752	0.570488	0.835299	0.946506	0.743370	1.000000	0.343357	0.022780	0.539726	0.545537	0.807958	0.6
yellow_cards	0.565963	0.592223	0.097151	0.093761	0.209349	0.243912	0.249957	0.343357	1.000000	0.165064	0.093270	0.089065	0.401884	0.5
red_cards	0.160326	0.186333	0.053679	0.048815	-0.021444	0.006284	0.073932	0.022780	0.165064	1.000000	0.055542	0.047354	0.104005	0.1
npg	0.437110	0.392631	0.971591	0.894286	0.587316	0.585152	0.852989	0.539726	0.093270	0.055542	1.000000	0.913496	0.720978	0.2
npxG	0.465546	0.408231	0.905710	0.979218	0.615503	0.611100	0.901386	0.545537	0.089065	0.047354	0.913496	1.000000	0.763481	0.2
xGChain	0.726598	0.703801	0.727953	0.763909	0.752587	0.814487	0.843152	0.807958	0.401884	0.104005	0.720978	0.763481	1.000000	0.8
xGBuildup	0.697196	0.731377	0.290990	0.282746	0.473254	0.547983	0.448197	0.618754	0.562467	0.167660	0.284135	0.273090	0.802073	1.0

Analisa dengan groupby

- Method `groupby` memungkinkan analisa dilakukan secara per kelompok nilai atribut tertentu. Misal: rerata dan simpangan baku gol per tim.

```
In [30]: df.groupby('team_title')['goals'].std()
```

```
Out[30]: team_title
Arsenal                          3.352381
Arsenal,Brighton                      NaN
Arsenal,Newcastle United              NaN
Arsenal,West Bromwich Albion          NaN
Aston Villa                      3.696489
Aston Villa,Chelsea                   NaN
Brighton                         2.158703
Burnley                          2.475210
Chelsea                          2.350177
Chelsea,Fulham                        NaN
Crystal Palace                   2.901461
Everton                          3.467727
Everton,Southampton                   NaN
Fulham                           1.439175
Leeds                            4.153193
Leicester                        4.020602
Liverpool                        4.931439
Liverpool,Southampton                 NaN
Manchester City                  3.867132
Manchester United                4.317855
```

```
In [29]: df.groupby('team_title')['goals'].mean()
```

```
Out[29]: team_title
Arsenal                          1.961538
Arsenal,Brighton                 0.000000
Arsenal,Newcastle United         8.000000
Arsenal,West Bromwich Albion     0.000000
Aston Villa                      2.130435
Aston Villa,Chelsea              3.000000
Brighton                         1.500000
Burnley                          1.280000
Chelsea                          2.240000
Chelsea,Fulham                   1.000000
Crystal Palace                   1.625000
Everton                          1.607143
Everton,Southampton              3.000000
Fulham                           0.925926
Leeds                            2.608696
Leicester                        2.370370
Liverpool                        2.370370
Liverpool,Southampton            3.000000
Manchester City                  3.208333
Manchester United                2.518519
```

Quiz / Games

- Lihat berkas Jupyter Notebook terkait.

#JADIJAGOANDIGITAL

TERIMA KASIH

digitalent.kominfo

DTS_kominfo

digitalent.kominfo

digital talent scholarship