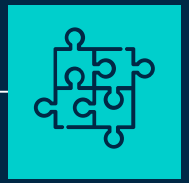
Peningkatan Kualitas Basis Data Pembangunan Jawa Tengah dengan Data Preparation

Pentingnya Kelengkapan Data untuk menghasilkan Analisa

Rismiyati, B.Eng, M.Cs

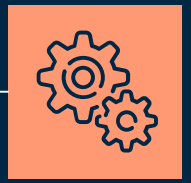
OVERVIEW PRESENTASI



01

Pentingnya Data

Kenapa data penting untuk pengambilan kebijakan dan bagaimana fakta yang terjadi



02

Proses Analisa Data

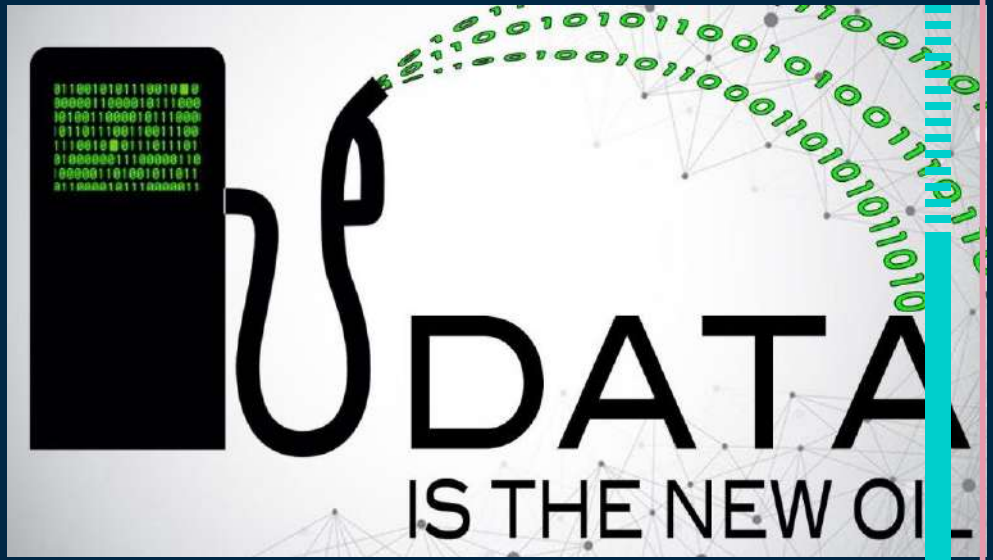
Proses dalam Data Science, terutama preprocessing

03

Pentingnya Kualitas Basis Data

Kualitas Basis Data dan kaitannya dengan analisa data dan Satu Data

PENTINGNYA DATA



Data mengubah dunia, karena banyak hal dapat dilakukan berdasarkan hasil eksploitasi data. Selain itu, fakta bahwa data tidak dapat habis dan dapat digunakan kembali membuat sumber daya ini bahkan lebih berharga daripada minyak.

Fakta

Baru sedikit perusahaan yang memanfaatkan data!!

Itu termasuk perusahaan dan lembaga pemerintahan di dunia. "Hanya 12% organisasi yang kami survei secara global mengerti dan mendapatkan *insight* dari data," kata Head of Presales Dell Technologies Indonesia Fitra Suryanto dalam Katadata Forum Virtual Series bertajuk 'Mengantisipasi dan Memanfaatkan Ledakan Data', Kamis (22/7).

- https://katadata.co.id/desysetyowati/digital/60f9215336d94/data-jadi-new-oil-tapi-baru-12-perusahaan-yang-memanfaatkan

Kenapa?

konsolidasi data

pemanfaatan data membutuhkan biaya besar

ego sektoral, terutama bagi lembaga pemerintahan

kesiapan infrastruktur digital

risiko keamanan data

PERATURAN GUBERNUR JAWA TENGAH

NOMOR 6 TAHUN 2022

TENTANG

SATU DATA JAWA TENGAH

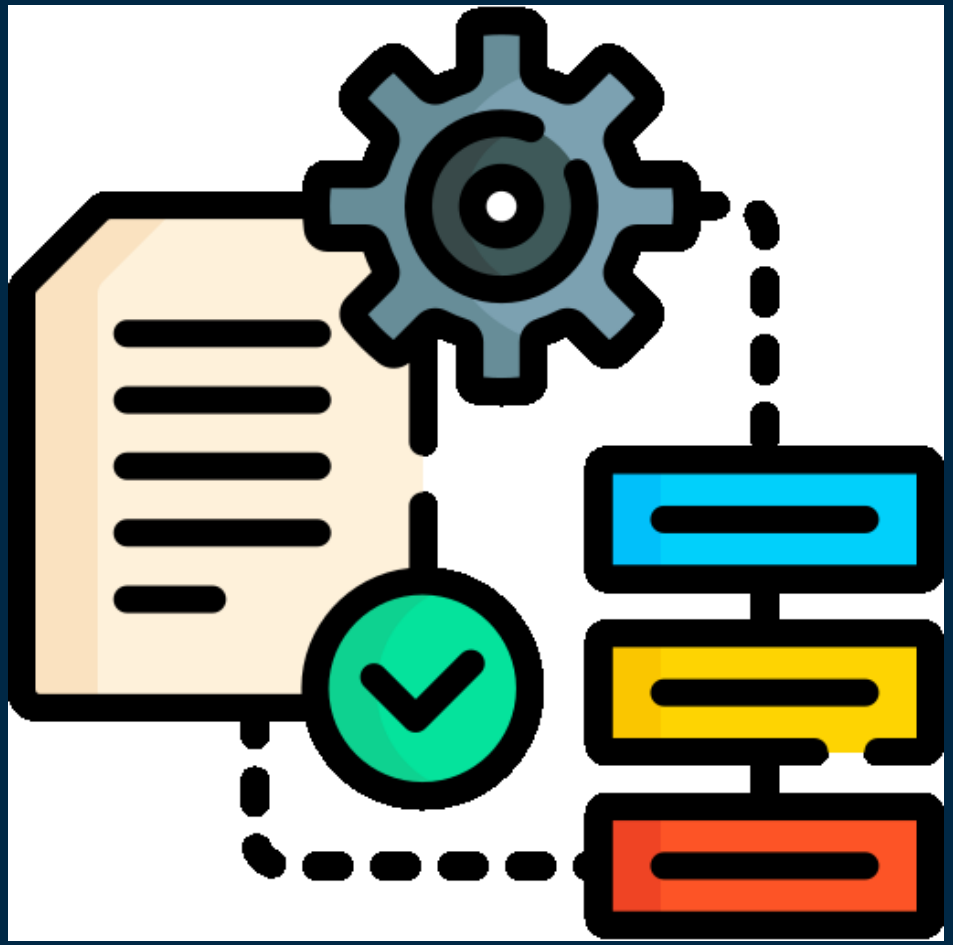
DENGAN RAHMAT TUHAN YANG MAHA ESA

GUBERNUR JAWA TENGAH,

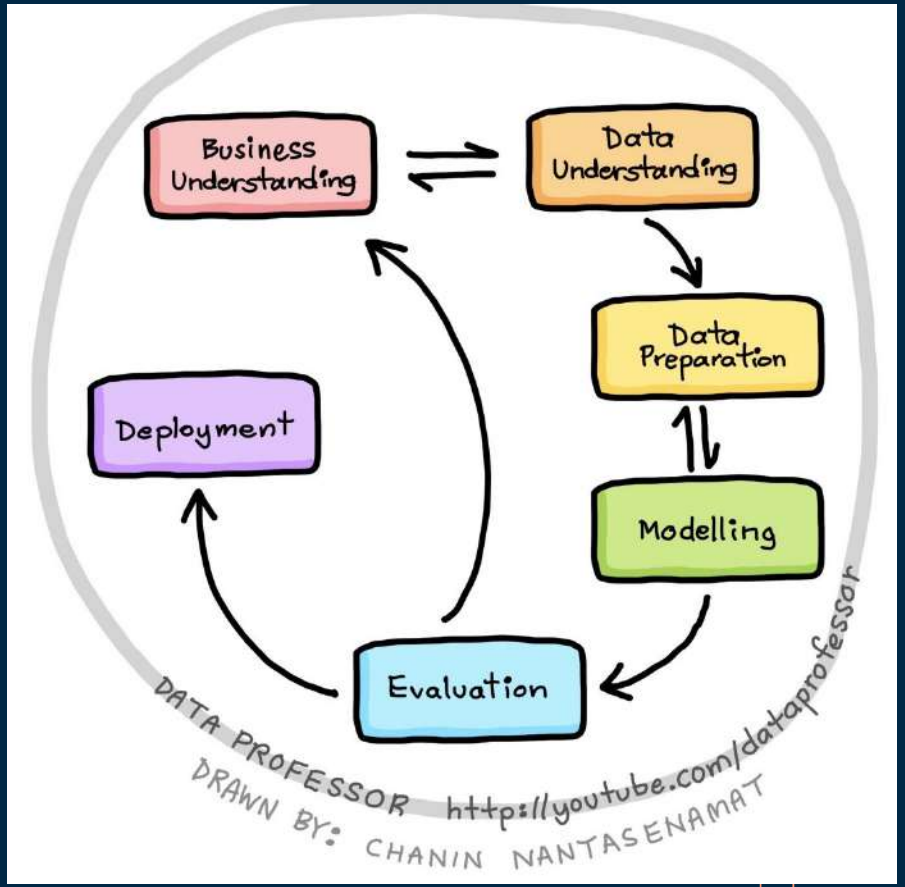
12. Satu Data Jawa Tengah adalah kebijakan tata kelola Data Pemerintah Provinsi Jawa Tengah untuk menghasilkan Data yang akurat, mutakhir, terpadu, dan dapat dipertanggungjawabkan, serta mudah diakses dan dibagipakaikan antar Instansi Pusat dan Instansi Daerah melalui pemenuhan Standar Data, Metadata, Interoperabilitas Data dan menggunakan Kode Referensi dan Data Induk.

Proses Analisa Data

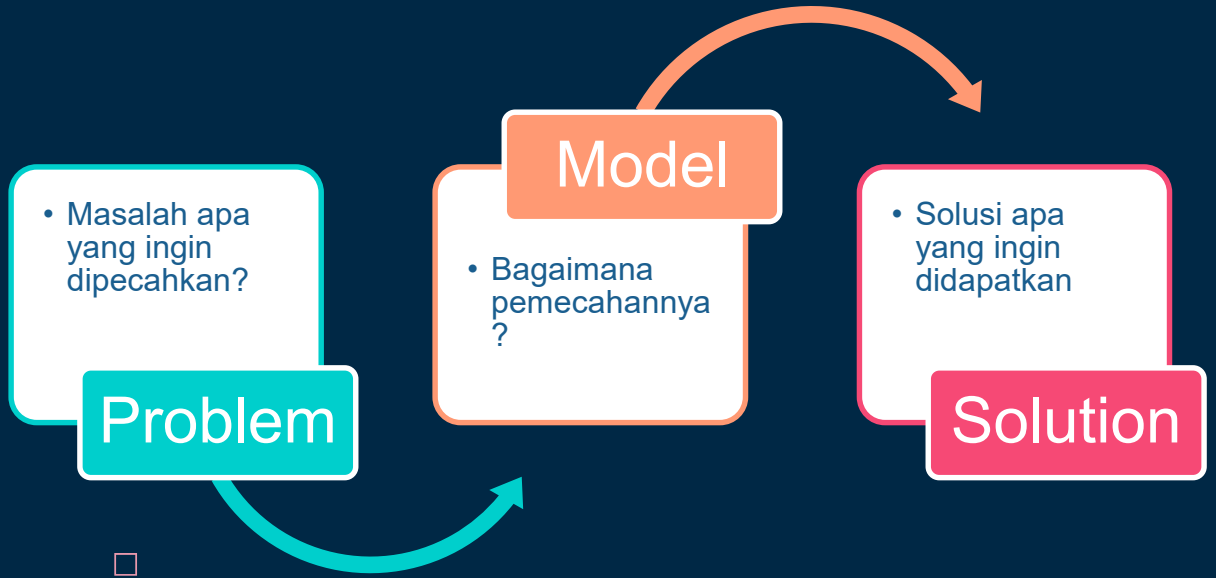
Bagaimana Data Sains dapat digunakan pada data yang tersedia



Langkah Data Science



DS Proses 1. Bussiness Understanding/Memahami Tujuan



Contoh: Data pelanggaran lalu lintas

- Apakah pria atau wanita lebih sering ngebut?
- Apakah jenis kelamin memengaruhi siapa yang diperiksa selama perhentian?
- Tahun berapa yang paling sedikit razia?
- Bagaimana aktivitas obat berubah menurut waktu?
- Prediksi ke depan pelanggaran apa yang berkaitan dengan aktifitas obat?

stop_date	stop_time	county_name	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw	violation	search_conducted	search_type	stop_outcome	is_arrested	stop_duration	drugs_related_stop
2005-01-02	01:55	NaN	M	1985.0	20.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
2005-01-18	08:15	NaN	M	1965.0	40.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
2005-01-23	23:15	NaN	M	1972.0	33.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False
2005-02-20	17:15	NaN	M	1986.0	19.0	White	Call for Service	Other	False	NaN	Arrest Driver	True	16-30 Min	False
2005-03-14	10:00	NaN	F	1984.0	21.0	White	Speeding	Speeding	False	NaN	Citation	False	0-15 Min	False

DS Proses #2. Data Understanding

- Fase ini memungkinkan kita untuk membiasakan diri dengan data
- Dilakukan dengan analisis data eksplorasi.
- Eksplorasi data awal tersebut memungkinkan kita untuk mengetahui subset data mana yang akan digunakan untuk pemodelan lebih lanjut serta membantu dalam menghasilkan hipotesis untuk dijelajahi.

Eksploratory Data Analysis

```
Data columns (total 15 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   stop_date           91741 non-null  object
 1   stop_time           91741 non-null  object
 2   county_name         0 non-null      float64
 3   driver_gender       86406 non-null  object
 4   driver_age_raw      86414 non-null  float64
 5   driver_age          86120 non-null  float64
 6   driver_race         86408 non-null  object
 7   violation_raw       86408 non-null  object
 8   violation           86408 non-null  object
 9   search_conducted    91741 non-null  bool
 10  search_type         3196 non-null   object
 11  stop_outcome        86408 non-null  object
 12  is_arrested         86408 non-null  object
 13  stop_duration       86408 non-null  object
 14  drugs_related_stop  91741 non-null  bool
```

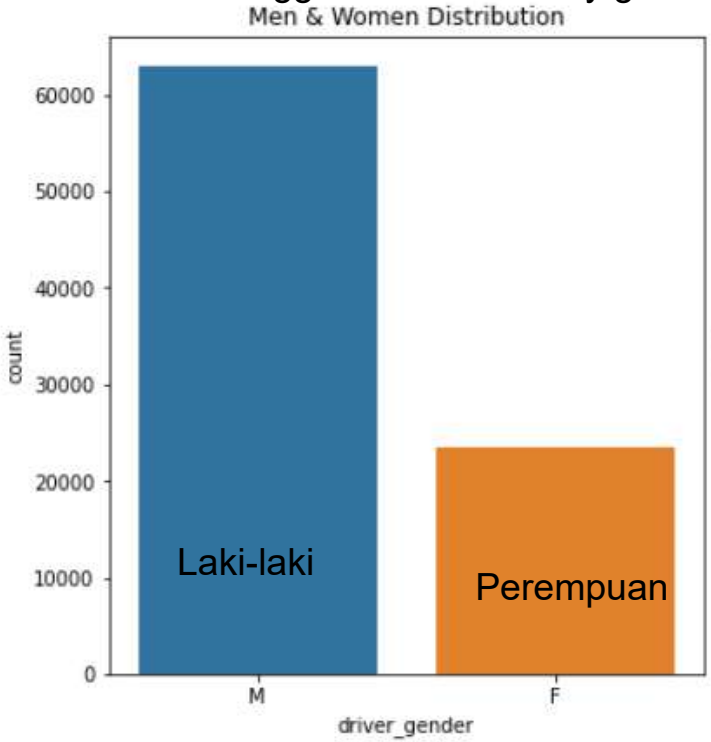
```
df.isnull().sum()
```

```
stop_date              0
stop_time              0
county_name        91741
driver_gender       5335
driver_age_raw      5327
driver_age          5621
driver_race         5333
violation_raw       5333
violation           5333
search_conducted       0
search_type        88545
stop_outcome        5333
is_arrested         5333
stop_duration       5333
drugs_related_stop     0
dtype: int64
```

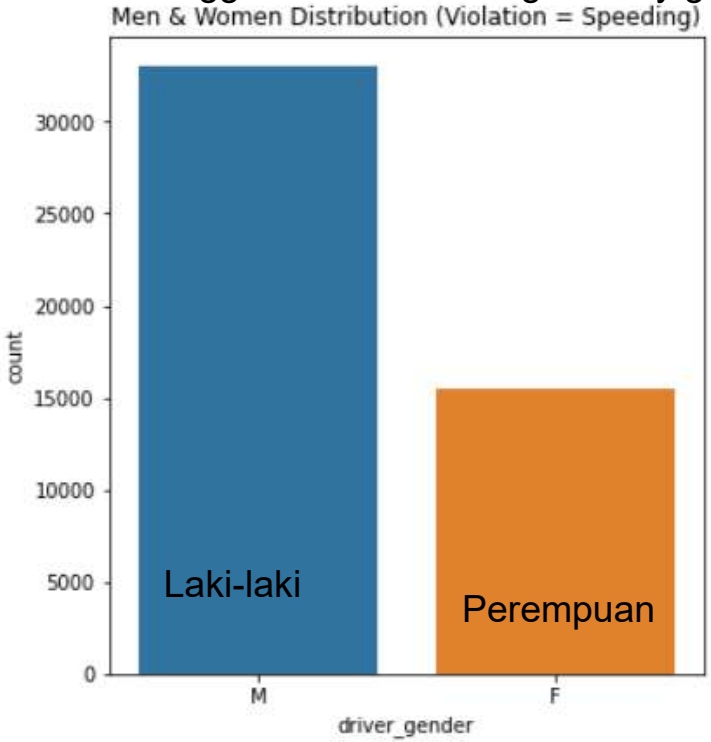
	county_name	driver_age_raw	driver_age
count	0.0	86414.000000	86120.000000
mean	NaN	1970.491228	34.011333
std	NaN	110.914909	12.738564
min	NaN	0.000000	15.000000
25%	NaN	1967.000000	23.000000
50%	NaN	1980.000000	31.000000
75%	NaN	1987.000000	43.000000
max	NaN	8801.000000	99.000000

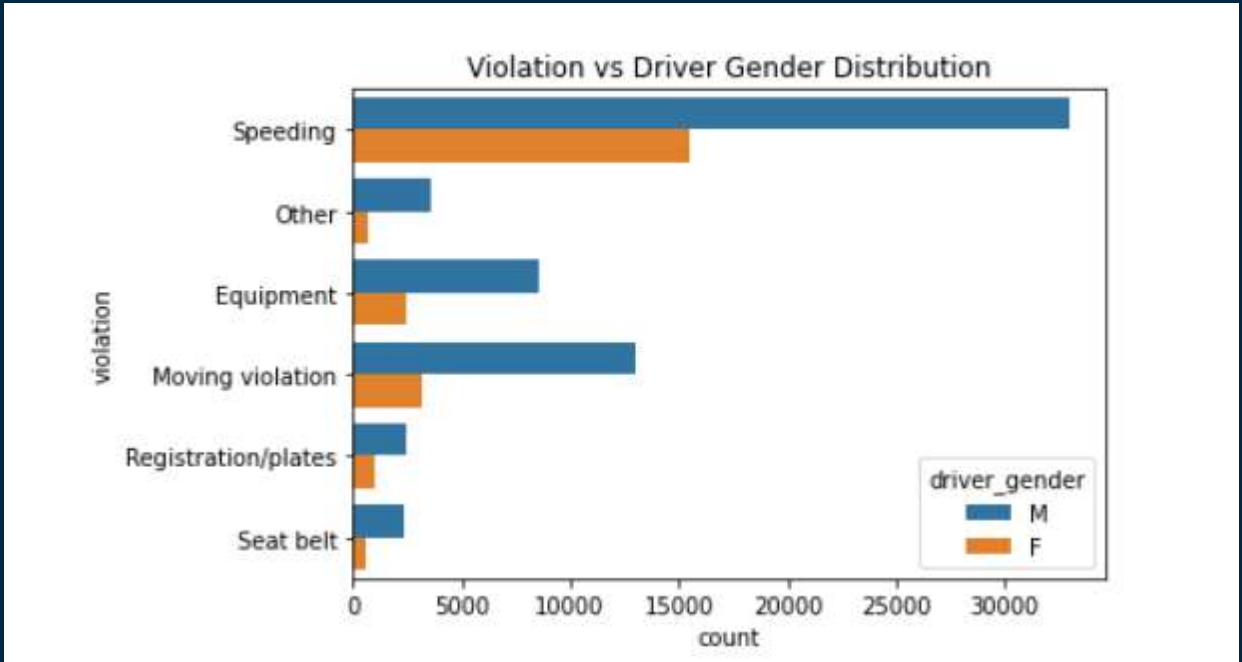
NaN, singkatan dari bukan angka, adalah anggota dari tipe data numerik yang dapat diartikan sebagai nilai yang tidak terdefinisi atau tidak terwakili, terutama dalam aritmatika floating-point

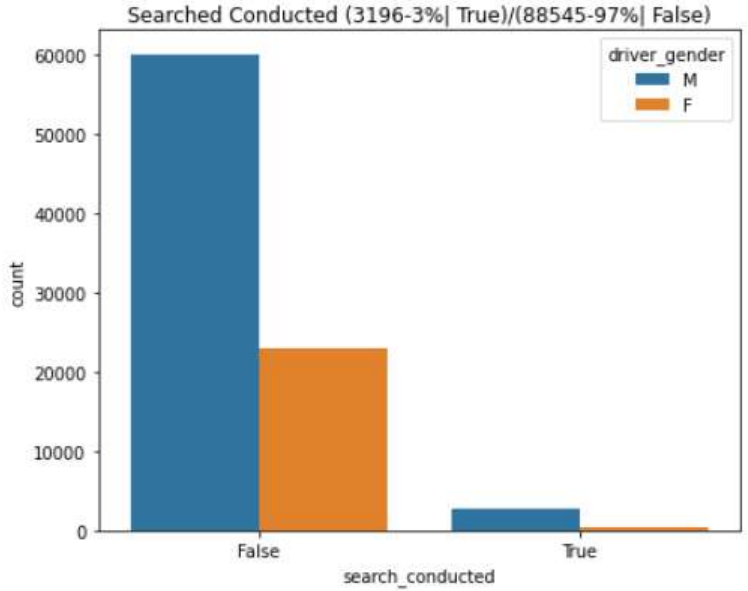
Pelanggaran lalu lintas by gender

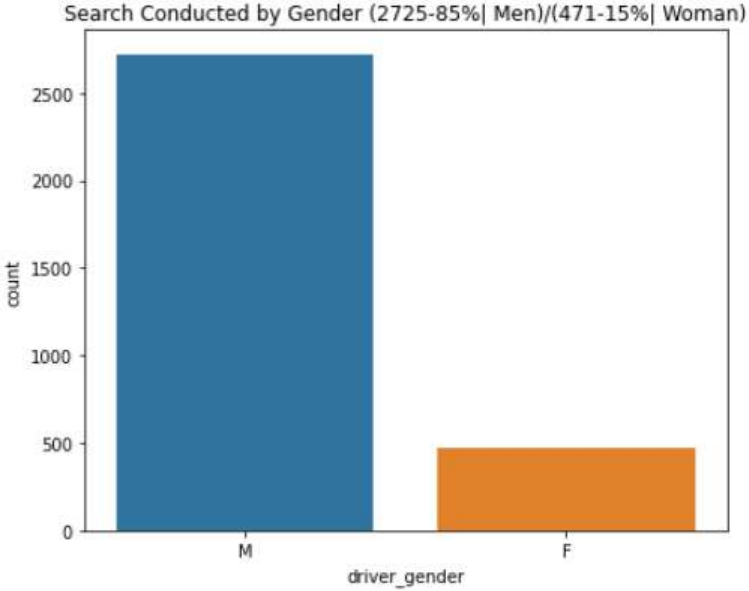


Pelanggaran lalu lintas: ngebut: by gender

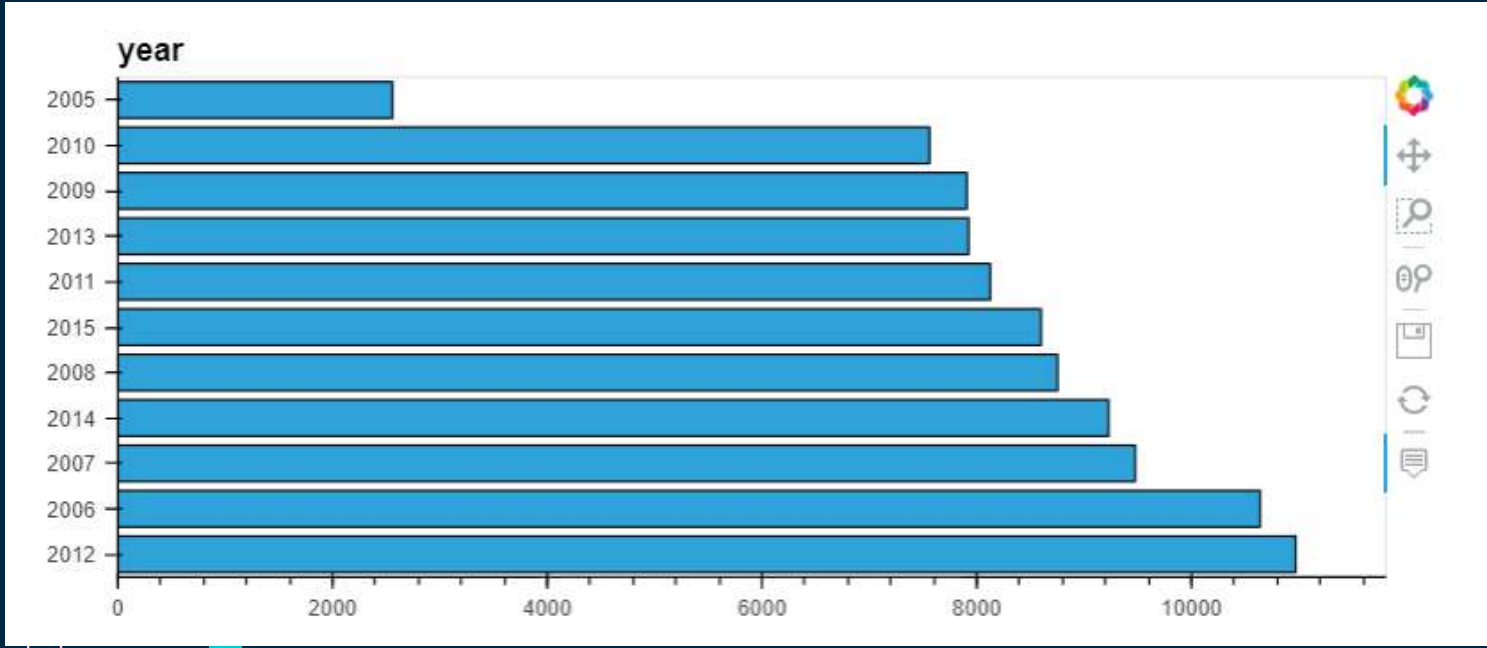



Violation vs Driver Gender Distribution
Speeding
Other
Equipment
Moving violation
Registration/plates
Seat belt
violation
driver_gender
M
F
0
5000
10000
15000
20000
25000
30000
count

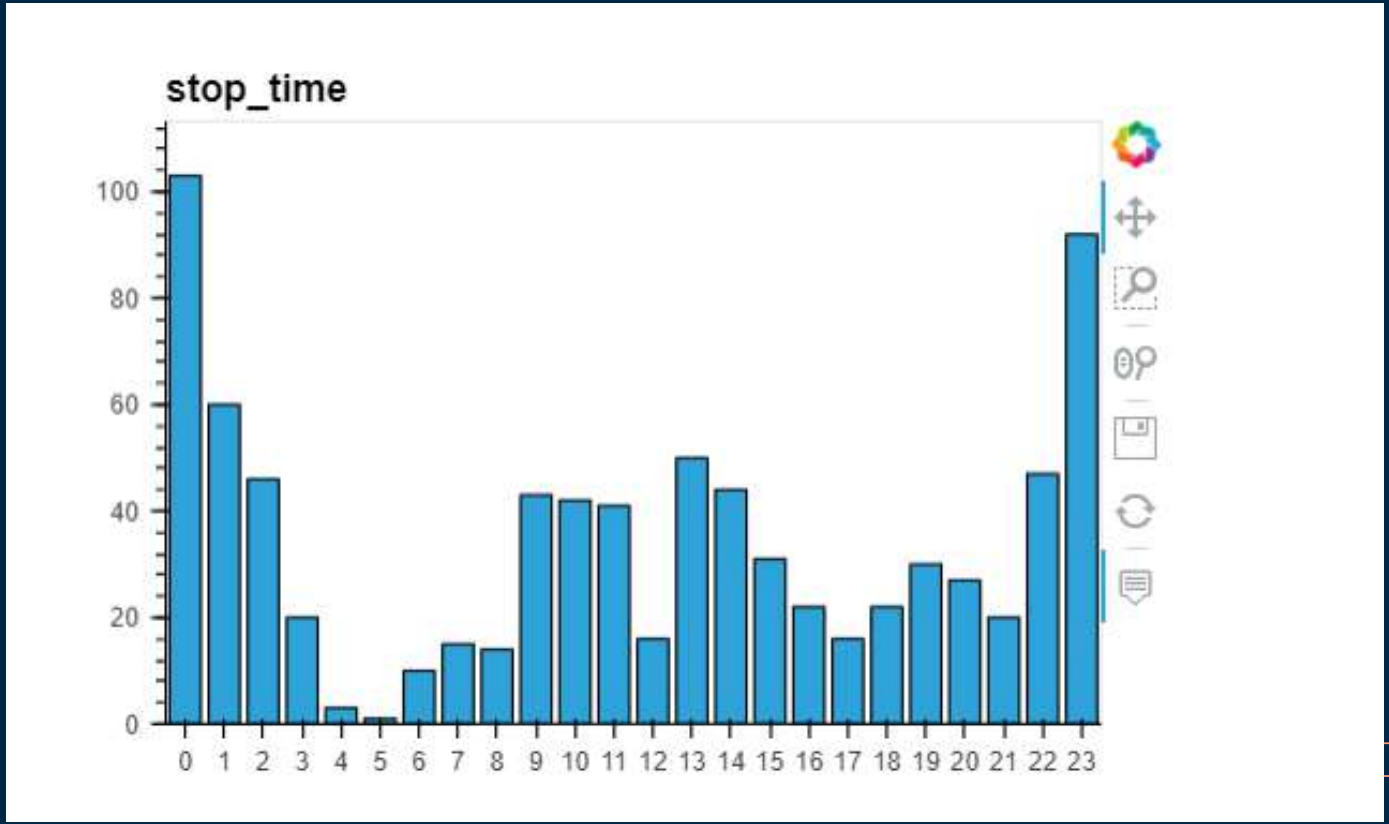

Searched Conducted (3196-3%| True)/(88545-97%| False)
driver_gender
M
F
60000
50000
40000
30000
20000
10000
0
count
False
True
search_conducted


Search Conducted by Gender (2725-85%| Men)/(471-15%| Woman)
2500
2000
1500
1000
500
0
count
M
F
driver_gender

Tahun berapa yang paling sedikit razia?



Pemeriksaan yang berkaitan dengan obat-obatan terlarang

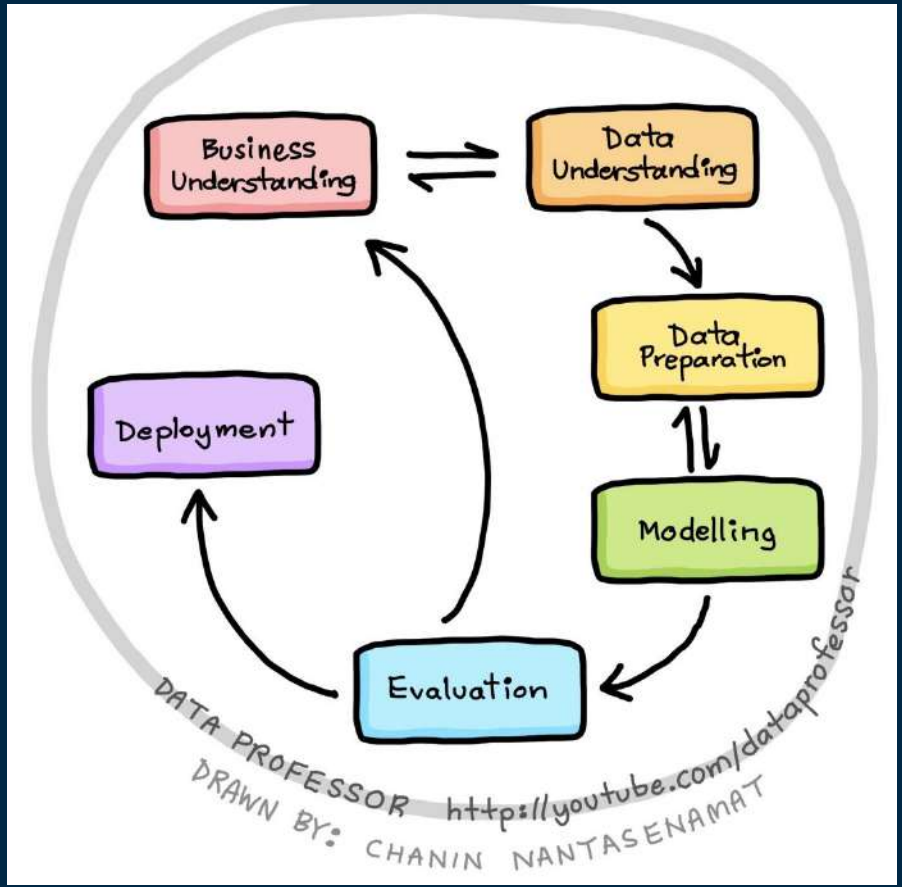


DS Proses #3. Data Preparation

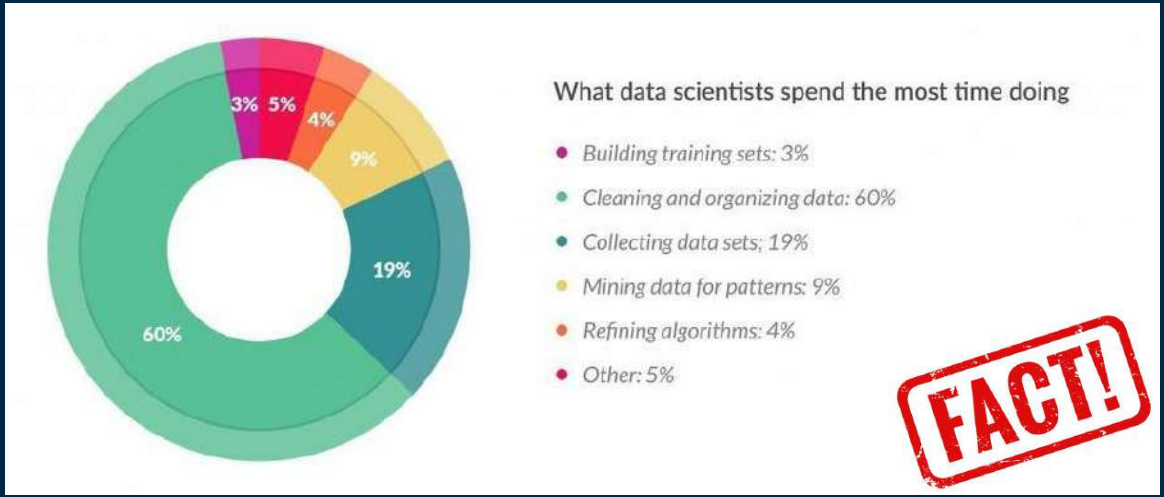
"With data, failing to prepare is preparing to fail"

Data Preparation

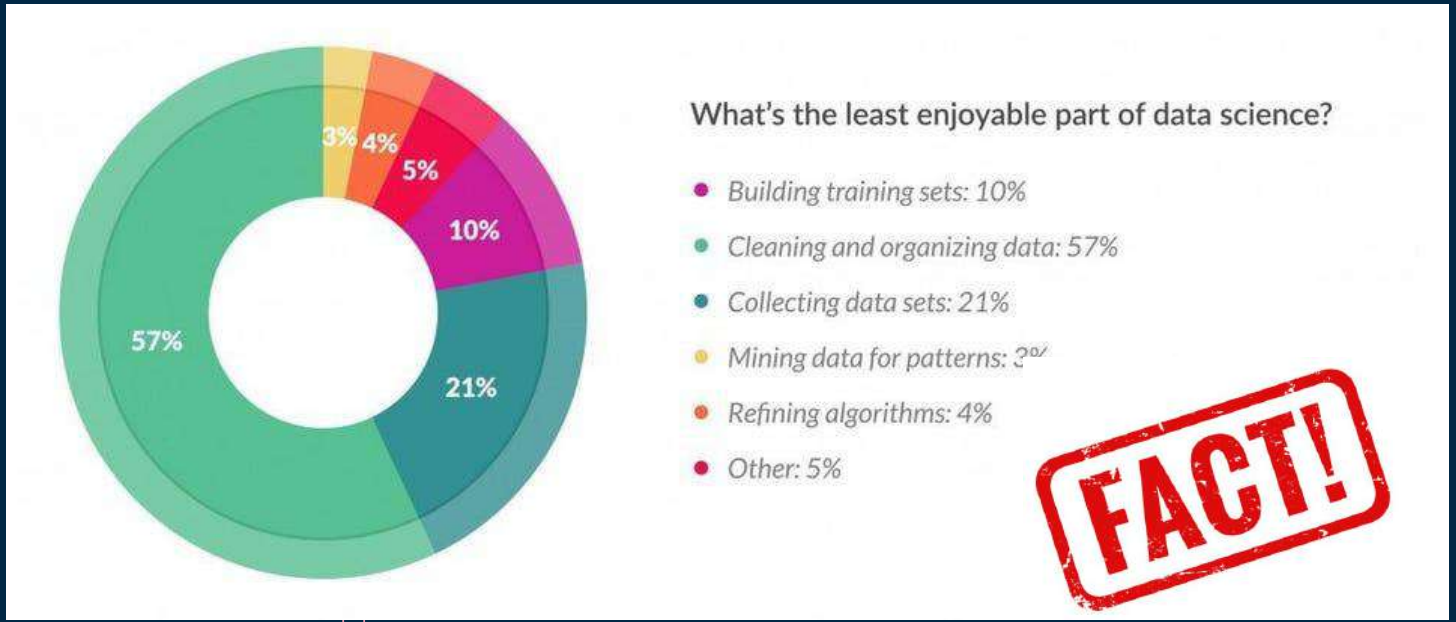
- Data preparation serangkaian proses pembersihan (*cleaning*) dan transformasi *raw data* yang dilakukan sebelum proses analisis/ modelling/ reporting untuk menjadikan data clean, akurat, lengkap dan realiable.



Fakta tentang Data Preparation



Fakta tentang Data Preparation



Mengapa Data Preparation Penting?

- Data didapat dari **berbagai macam sumber** yang umumnya didesain dalam format untuk menjalankan fungsi tertentu dalam aplikasi. Hal ini dapat berbeda dengan apa yang diperlukan untuk proses analisis.
- Beberapa algortima machine learning memerlukan data dalam format tertentu, misalnya format numerik.
- Data harus bersih, akurat, lengkap dan *well-labelled* sebelum proses analisis dilakukan.
- Kegagalan utama dalam analisis data maupun project machine learning disebabkan oleh data yang tidak berkualitas, bukan dari algoritma/ code maupun vendornya.

Tujuan Data Preparation

- Meningkatkan keseluruhan kualitas data dan proses analisisnya.
- Mempercepat proses analisis pengambilan keputusan
- Menggabungkan sejumlah dataset yang terpisah ke dalam satu format.
- Mendapatkan hasil analisis data yang lebih berkualitas dan akurat.

Task in Data Preparation

- **Data Cleaning**
  - Mengidentifikasi dan memperbaiki error yang ada di dalam dataset yang dapat memberikan dampak negatif pada proses analisis berikutnya, contohnya: menghapus duplikasi data dan data yang tidak relevan, menangani missing dan outlier, dsb.
- **Transformasi Data**
  - Transformasi data ke format yang diperlukan sehingga dapat lebih dipahami dan dianalisis dengan mudah.
- **Reduksi Dimensi**
  - Mengurangi dimensi data yang terlalu besar untuk meningkatkan efisiensi dalam proses analisis/ pemodelan dan mendapatkan hasil analisis yang lebih baik.

Task in Data Preparation (lanj.)

- Proses yang dilakukan dalam data preparation dapat berbeda-beda, bergantung pada:
  - Tujuan analisis data
  - Keahlian/ expertise pengguna data
  - Pertanyaan apa yang ingin di jawab dari data tersebut.

Aspek dalam Data Preparation

- *Data completeness*
  - Apakah data yang ada cukup untuk merepresentasikan pattern of interest? Adakah tambahan data yang diperlukan dari sumber lainnya?
- *Data correctness*
  - Apakah data akurat? Apakah data sudah clean, standardize, dll?
- *Data quantity*
  - Apakah jumlah data yang tersedia sudah mencukupi?
- *Data usability*
  - Apakah data sudah siap digunakan untuk proses analisis lebih lanjut?

Data Cleaning

- Mengidentifikasi dan memperbaiki error yang ada di dalam dataset yang dapat memberikan dampak negatif pada proses analisis berikutnya.
- Task dalam data cleaning
    1. Menghilangkan data yang tidak relevan, termasuk duplikasi data
    2. Memperbaiki *structural errors*
    3. Menangani outlier
    4. Menangani missing data
    5. Validasi data dan QA (Quality Assurance)

In [8]:

```
for column in df.columns:
    missing = df[column].isna().sum() / df.shape[0]
    print(f"{column:{20}}: ==============> {missing * 100:.2f}%")
```

```
stop_date           : ==============> 0.00%
stop_time           : ==============> 0.00%
county_name         : ==============> 100.00%
driver_gender       : ==============> 5.82%
driver_age_raw      : ==============> 5.81%
driver_age          : ==============> 6.13%
driver_race         : ==============> 5.81%
violation_raw       : ==============> 5.81%
violation           : ==============> 5.81%
search_conducted    : ==============> 0.00%
search_type         : ==============> 96.52%
stop_outcome        : ==============> 5.81%
is_arrested         : ==============> 5.81%
stop_duration       : ==============> 5.81%
drugs_related_stop  : ==============> 0.00%
```

In [9]:

```
df.dropna(axis=1, how='all').shape
```

Data Cleaning #1 Menghilangkan data yang tidak relevan

- Mengapus data yang tidak sesuai dengan permasalahan yang akan dianalisis.
    - Contoh: sebuah perusahaan ingin menganalisis profil customer usia muda, maka data mengenai customer yang telah parobaya/tua dapat dihilangkan.
- Menghapus kolom/fitur yang tidak diperlukan: Kolom yang berisi nilai unik pada setiap barisnya (seperti ID) atau kolom yang memiliki banyak sekali missing data.
- Identifikasi dan menghapus kolom yang hanya berisi single-value
- Identifikasi dan menghapus kolom yang memiliki low-variance
- Identifikasi dan menghapus duplikasi baris data

Data Cleaning #2 Memperbaiki Structural Errors

- Memperbaiki kesalahan pada naming convention, typo, incorrect capitalization, atau inkonsistensi pada kategori atau pelabelan data.
- Contoh:
  - Perempuan, P, Wanita, W ☐ harusnya diberikan nilai yang sama untuk menyatakan jenis kelamin yang sama.
  - Semarang, SMG, Smg, Smarang ☐ harusnya diberikan nilai yang sama untuk menyatakan kota yang sama.

```
0-15 Min     69543
16-30 Min    13635
NaN           5333
30+ Min       3228
2                1
1                1
Name: stop_duration, dtype: int64
```

```
# ri.stop_duration.replace(['1', '2'], value=np
df.loc[(df.stop_duration == '1')| (df.stop_dur
```

```
df.stop_duration.value_counts(dropna=False)
```

```
0-15 Min     69543
16-30 Min    13635
NaN           5335
30+ Min       3228
Name: stop_duration, dtype: int64
```

2. Memperbaiki Structural Errors (lanj.)

```
1 from pandas import read_csv
2 # load dataset
3 df = read_csv('/content/drive/MyDrive/python/ice/dataset/BL-Flickr-Images-Book.csv')
4 df.head(5)
```

	Identifier	Edition Statement	Place of Publication	Date of Publication	Publisher	Title	Author	Contributors	Corporate Author	Corporate Contributors	Former owner	Engraver	Issuance type
0	206	NaN	London	1879 [1878]	S. Tinsley & Co.	Walter Forbes. [A novel.] By A. A	A. A.	FORBES, Walter.	NaN	NaN	NaN	NaN	monographic
1	216	NaN	London; Virtue & Yorston	1868	Virtue & Co.	All for Greed. [A novel. The dedication signed...	A., A. A.	BLAZE DE BURY, Marie Pauline Rose - Baroness	NaN	NaN	NaN	NaN	monographic
2	218	NaN	London	1869	Bradbury, Evans & Co.	Love the Avenger. By the author of "All for Gr...	A., A. A.	BLAZE DE BURY, Marie Pauline Rose - Baroness	NaN	NaN	NaN	NaN	monographic
3	472	NaN	London	1851	James Darling	Welsh Sketches, chiefly ecclesiastical, to the...	A., E. S.	Appleyard, Ernest Silvanus.	NaN	NaN	NaN	NaN	monographic

Data Cleaning #3 Menangani Outlier

- Outliers (pencilan) adalah sampel yang sangat jauh dari mainstream data.
- Outlier umumnya jarang, berbeda dan tidak sesuai dalam beberapa hal.
- Tidak ada cara yang pasti untuk mendefinisikan dan mengidentifikasi outlier secara umum karena sifatnya sangat spesifik pada setiap dataset.
- Domain expert dapat dilibatkan untuk menginterpretasikan dan menentukan apakah sebuah sampel merupakan outlier atau tidak.
- Outlier dapat disebabkan oleh:
  - Data entry error, measurement error, sampling error, atau experimental error.
  - Data corruption atau data processing error.
  - Natural (true outlier observation)

Data Cleaning #4 Menangani Missing Value

- Missing value : Baris data dalam dataset dimana satu atau beberapa nilai kolomnya tidak tersedia.
- Missing value dapat ditandai dengan:
  - Empty value, empty string, question mark (?), NULL, undefined (N/A), angka 0, NaN (Not a Number) atau nilai di luar jangkauan (out-of-range entries)
- Kemungkinan (*likelihood*) missing value semakin meningkat seiring bertambahnya ukuran dataset.
- Penyebab missing value: masalah spesifik pada problem domain, observasi yang tidak tercatat, data corruption, data unavailability, malfunctioning measurement equipment, dsb.

Data Cleaning #5 Validasi Data

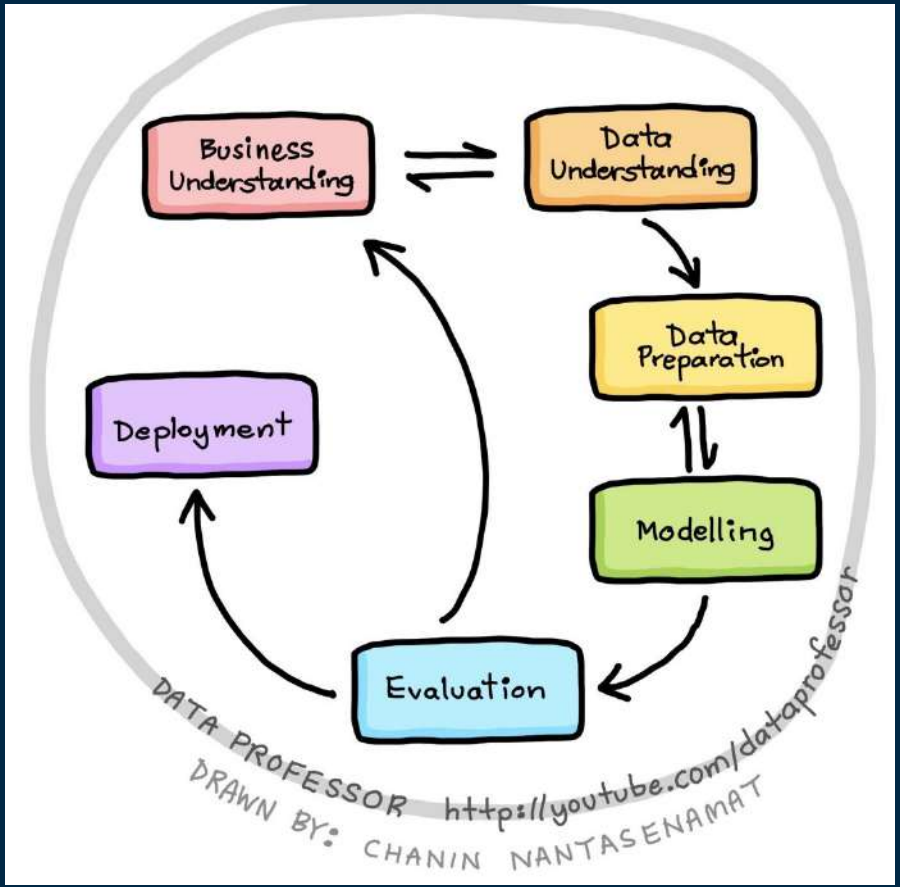
- Beberapa pertanyaan yang perlu dijawab untuk memastikan validasi data:
  - Apakah datanya masuk akal (make sense)?
  - Apakah data mengikuti aturan yang sesuai pada domainnya?
  - Apakah data tersebut membuktikan atau menyangkal working theory, atau membawa insight yang baru?
  - Dapatkah tren dalam data ditemukan untuk membantu membentuk teori berikutnya?
  - Jika tidak, apakah itu karena masalah kualitas data?

Data Cleaning #5 Validasi Data

- Berikut ini adalah 5 karakteristik dari kualitas data:
  - *Validity* ☐ sejauh mana data sesuai dengan aturan atau batasan bisnis yang ditentukan.
  - *Accuracy* ☐ pastikan data mendekati nilai sebenarnya.
  - *Completeness* ☐ sejauh mana semua data yang diperlukan diketahui.
  - *Consistency* ☐ pastikan data konsisten dalam set data yang sama dan/atau di beberapa set data.
  - *Uniformity* ☐ sejauh mana data ditentukan dengan menggunakan satuan ukuran yang sama.

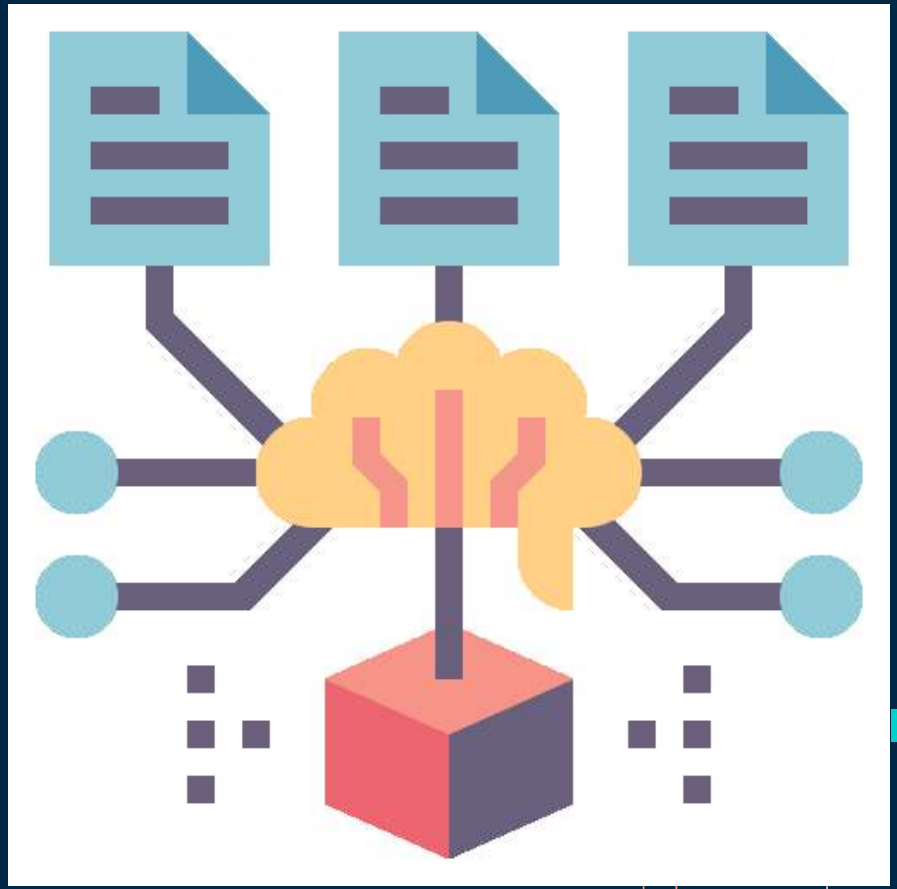
4. Pemodelan

- data hasil *Preprocessin*g digunakan untuk membangun model di mana algoritma pembelajaran digunakan untuk melakukan analisis multivariat.
- Contoh: mencari model untuk melakukan prediksi jumlah panen padi di Kabupaten Kendal pada triwulan ketiga 2022

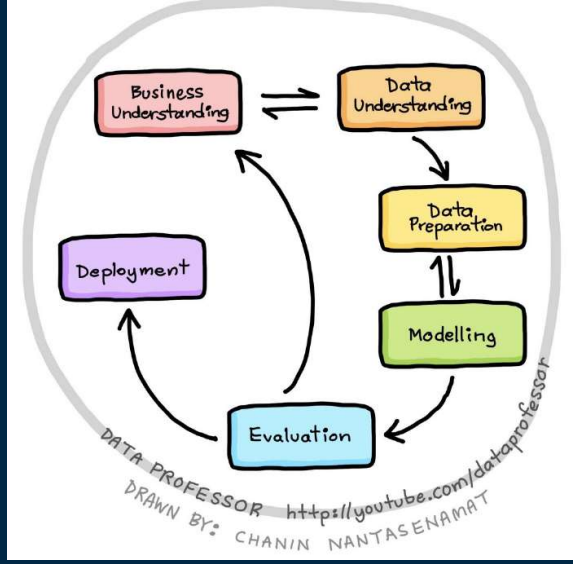


4. Pemodelan

- Contoh: Menggunakan algoritma Jaringan syaraf tiruan untuk menentukan apakah suatu pelanggaran merupakan pelanggaran berkaitan dengan obat-obatan



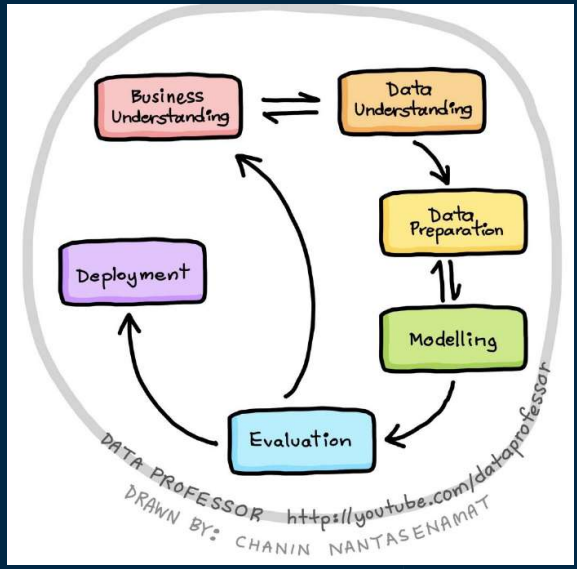
5. Evaluasi dan Deployment



- Dalam melakukan 4 langkah sebelumnya, penting untuk mengevaluasi hasil yang didapat dan meninjau proses yang dilakukan sejauh ini untuk menentukan apakah tujuan bisnis yang ditetapkan semula terpenuhi atau tidak.
- Jika dianggap tepat, beberapa langkah mungkin perlu dilakukan lagi.
- Setelah dirasa hasil dan prosesnya memuaskan maka kita siap untuk pindah ke deployment:
  - simple report
  - API that can be accessed via programmatic calls
  - a web application

Recap

- Dalam pemanfaatan *Data Science* tidak berarti semua langkah harus dilakukan
- Bergantung pada tujuan yang ingin dicapai



Pentingnya Kualitas Basis Data

Kualitas Basis Data dan kaitannya dengan analisa data dan Satu Data

Basis Data dan Kualitas Data

- Basis Data adalah **kumpulan data** yang dikelola sedemikian rupa berdasarkan ketentuan tertentu yang saling berhubungan sehingga mudah dalam pengelolaannya
- kualitas data adalah penilaian seberapa banyak data dapat digunakan dan sesuai dengan konteks penyajiannya.

Ukuran Kualitas Data

- **Konsistensi Data**: Pelanggaran aturan semantik yang ditentukan pada kumpulan data
- **Akurasi Data**: Data akurat ketika nilai data yang disimpan dalam database sesuai dengan nilai dunia nyata.
- **Keunikan Data:** Ukuran duplikasi yang tidak diinginkan yang ada di dalam atau di seluruh sistem untuk bidang, catatan, atau kumpulan data tertentu.
- **Kelengkapan Data**: Sejauh mana nilai-nilai hadir dalam pengumpulan data.
- **Ketepatan Waktu Data**: Sejauh mana usia data disesuaikan untuk tugas yang ada. [3]

Kenapa Kualitas Data Penting?

- kualitas data yang buruk menyebabkan **pelaporan yang tidak akurat** yang akan menghasilkan **keputusan yang tidak akurat** dan tentunya kerugian ekonomi

Bagaimana Meningkatkan Kualitas Data?

- Pelatihan pada tenaga input data
- Menerapkan proses peningkatan kualitas data/ Data preprocessing (Costly)

Kaitan *Data Science*, kualitas data dengan Satu Data

Menimbang :

a. bahwa untuk mewujudkan keterpaduan perencanaan, pelaksanaan, evaluasi, dan pengendalian pembangunan Daerah, perlu didukung dengan data yang akurat, mutakhir, terpadu, dapat dipertanggungjawabkan, mudah diakses, dibagipakaikan, dikelola secara seksama, terintegrasi dan berkelanjutan;

b. bahwa untuk memperoleh data yang akurat, mutakhir, terpadu, dapat dipertanggungjawabkan, mudah diakses dan dibagipakaikan, diperlukan perbaikan Tata Kelola Data yang dihasilkan oleh Pemerintah Daerah melalui penyelenggaraan Satu Data;

Analisa data

Kualitas data

(1) Satu Data Jawa Tengah dilakukan berdasarkan prinsip sebagai berikut:

a. Data yang dihasilkan oleh Produsen Data Daerah harus memenuhi Standar Data;

b. Data yang dihasilkan oleh Produsen Data Daerah harus dilengkapi dengan Metadata;

c. Data yang dihasilkan oleh Produsen Data Daerah harus memenuhi kaidah Interoperabilitas Data; dan

d. Data yang dihasilkan oleh Produsen Data Daerah harus menggunakan Kode Referensi dan/atau Data Induk.

Kaitan *Data Science*, kualitas data dengan Satu Data

(1) Standar data sebagaimana dimaksud dalam Pasal 7 ayat (1) huruf a merupakan standar yang mendasari data tertentu dan terdiri atas:
   a. konsep;
   b. definisi;
   c. klasifikasi;
   d. ukuran; dan
   e. satuan.

(2) Konsep sebagaimana dimaksud pada ayat (1) huruf a mengacu pada ide yang mendasari data dan tujuan data tersebut diproduksi.

(3) Definisi sebagaimana dimaksud pada ayat (1) huruf b mengacu pada penjelasan tentang data yang memberi batas atas atau secara jelas membedakan arti dan cakupan dari data tertentu dengan data yang lain.

(4) Klasifikasi sebagaimana dimaksud pada ayat (1) huruf c mengacu pada penggolongan secara sistematis ke dalam kelompok atau kategori dalam data berdasarkan kriteria yang telah disepakati atau dibakukan secara luas.

(5) Ukuran sebagaimana dimaksud pada ayat (1) huruf d mengacu pada unit yang digunakan dalam pengukuran jumlah, kadar, atau cakupan sesuatu.

(6) Satuan sebagaimana dimaksud pada ayat (1) huruf e merupakan jumlah tunggal tertentu dalam data yang digunakan sebagai standar untuk mengukur atau menakar sesuatu sebagai sebuah keseluruhan.

Kesimpulan

- Data adalah aset berharga untuk semua lembaga
- Proses Analisa dapat dilakukan dengan dimulai dari pemahaman masalah, pemahaman data, data preprocessing, modelling dan evaluation
- Preprocessing merupakan langkah awal dan mutlak, terutama jika data tidak lengkap atau tidak memenuhi standard
- Pentingnya data yang berkualitas untuk menghasilkan analisa yang akurat

Saran

- Agar dapat dimanfaatkan dengan baik, pelaksanaan **Satu Data** sebaiknya dimulai dengan disiplin input data.
- Data yang dimasukkan harus lengkap atribut, lengkap data, sesuai dengan format yang ada.
- Format baku beserta contoh penginputan dapat disediakan sebagai acuan masing-masing dinas untuk melakukan input data

rismiyati@live.undip.ac.id
2022