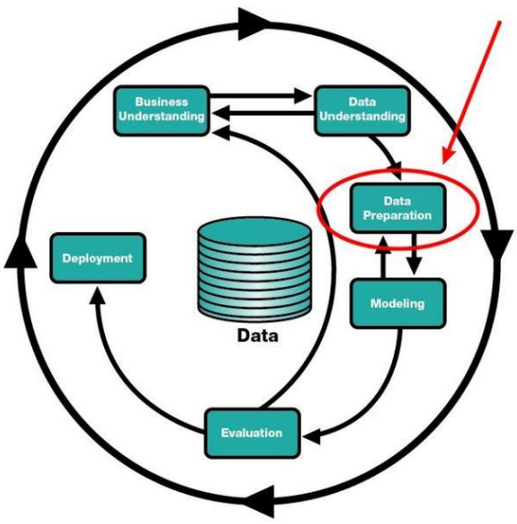
Data Preprocessing Using Python

Library Pandas pada Data Frame

Topik Content Chapter ini membahas tentang kegunaan Pandas, salah satunya untuk data processing (Filter, Cleansing, Aggregation dan manipulation). pandas library di analogikan sebagai function menu dalam excel (home,insert dll) sedangkan dataframe itu di analogikan sebagai raw data nya di dalam excel.



Saat ini teman-teman sedang berapa pada proses data preparation. dari proses ini teman-teman menyiapkan data yang sudah dikumpulkan dan diolah sehingga menjadi data yang siap di olah. untuk proses yang dilakukan pada data processing meliputi 1. data cleansing 2. data integration 3. data transformation 4. data reduction 5. feature engineering.

Proses detail yang dilakukan pada data processing seperti pada gambar dibawah ini :

Data Preparation Activities	What to do?	How to do?
Data Cleaning	Dealing with Missing Values/Features	• Ignore respective records having missing values or features • Substitute with dummy value, mean, mode, regressed values or values predicted by an algorithm
	Dealing with Duplicate values/ Redundant Data	• Deletion of duplicate or redundant records
	Dealing with Outliers and Noise	• Binning • Regression (smoothing or curve fitting) • Clustering (grouping values in cluster to identify and eliminate outliers)
	Dealing with Inconsistent/ Conflicting Data	• Use of domain expertise, business understanding, human discretion to correct the data
Data Integration (Integrate multiple sources)	Dealing with issues like Schema integration, entity identification and redundancy	• Joining data sets • Editing metadata to handle data inconsistencies like naming, type etc.
Data Transformation	• Generalization of data	• Concept hierarchy climbing to replace low level attributes with high level concepts or attributes (ex. 'Street' can be generalized to 'country')
	• Normalization/ Scaling of attribute values to a specified range	• Z-score method • Min-Max method • Decimal scaling
	• Aggregation	• Applying summary or aggregation operators to data (ex. Using daily sales to compute annual sales)
	• Feature Construction	• Add or replace with new features derived from existing ones
Data Reduction (Reducing data to make it easy to handle and produce similar analytical results)	• Dimensionality Reduction to eliminate insignificant features	• Feature Selection • Attribute Sampling • Heuristic Methods
	• Aggregation	• Use of aggregation techniques (as above)
	• Data Compression	• Reducing data size by using methods like wavelet transform, PCA etc.
	• Numerousity reduction to have smaller data representations	• Record Sampling, Clustering, Regression etc.
	• Generalization	• Concept hierarchy generation (as above)
Data Discretization (cont. features into discrete)	• Unsupervised (no label is used)	• Binning (equal-width and equal-depth)
	• Supervised (uses labels)	• Entropy-based
Feature Engineering	• Using or deriving the right features to improve accuracy of your analytical model	• Feature Selection • Validation & improvement of features • Brainstorming to create and test more features

Adapun basic operation pada dataframe sebagai berikut :

1. .info() -> melihat informasi dari dataframe
2. .shape -> melihat jumlah baris dan kolom
3. .columns -> melihat semua nama columns

4. .head(n) -> melihat jumlah baris n pertama
5. .tail(n) -> melihat jumlah baris n terakhir
6. .desribe() -> melihat statistik sederhana dari data
7. .sort_values() -> mengurutkan dataframe
8. .copy() -> mengcopy dataframe

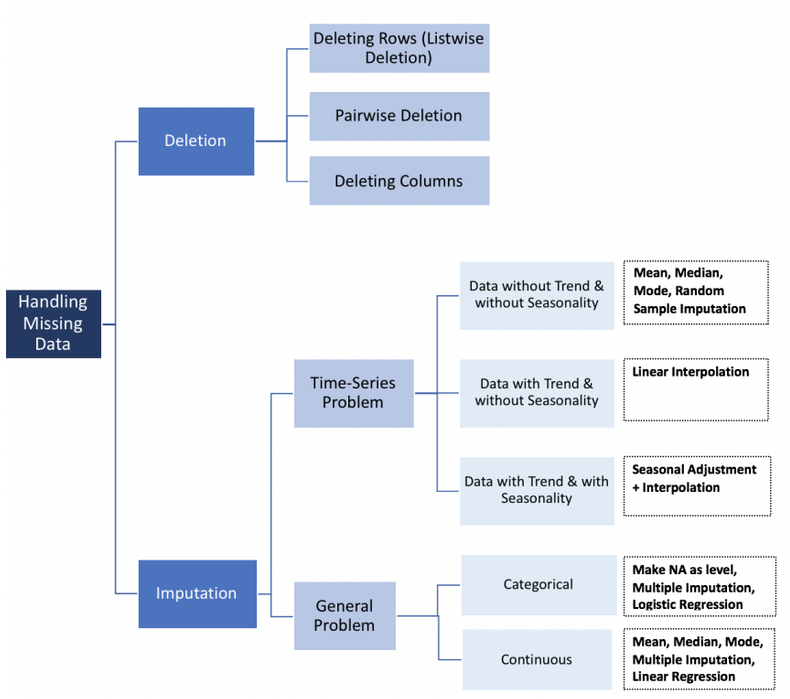
Basic Pandas Operation & Data Cleansing

setelah teman-teman memahami library dan kegunaan Pandas, maka langkah selanjutnya adalah memasukkan data tersebut kedalam environment python dengan menggunakan function import import pandas as pd dan sesuaikan dengan tipe file teman-teman yang ingin diimport.

Missing Value Treatment

Different types of missing values:

- Not Missing at Random: NMAR -> menemukan data case terbaik atau terburuk (bisa dengan menambahkan feature baru)
- Missing at Random: MAR -> treatment dengan mengisi sesuai dengan mean/median/modus
- Missing Completely at Random: MCAR -> treatment dengan drop data yang hilang rows/column

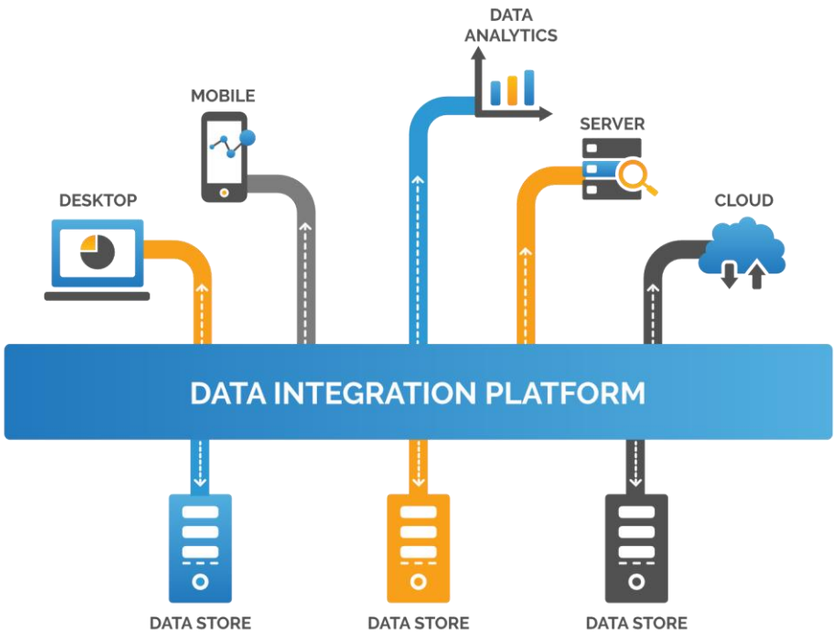


Data Integration

Fokus kita sekarang akan membahas Data Integration. data integration merupakan sebuah teknik untuk menggabungkan data dari berbagai sumber yang berbeda menjadi suatu informasi yang berharga.
Jenis Data Integration Menurut Xenonstack, terdapat dua jenis data integration, yaitu sebagai berikut :

1. Enterprise Data Integration (EDI) Jenis pertama dari data integration adalah Enterprise Data Integration (EDI) yaitu seperangkat teknologi yang dapat membantu memanipulasi kumpulan data.
   Data integration jenis ini melibatkan akuisisi data dari beragam sistem bisnis. Pengolahannya juga dilakukan dengan berbagai aktivitas manajemen dan laporan dari business intelligence.
2. Customer Data Integration (CDI) Selanjutnya adalah Customer Data Integration (CDI) adalah pengumpulan data dengan tujuan utamanya adalah memuaskan keinginan pelanggan.
   Jadi, jenis data integration yang satu ini adalah proses pengumpulan dan manipulasi data pelanggan secara terpadu sehingga bisa lebih mudah dianalisis.

Application Integration
Application integration melibatkan pemindahan data secara bolak-balik antar aplikasi untuk menjaganya tetap sinkron. Biasanya setiap aplikasi memiliki cara tertentu untuk menerima dan mengeluarkan data. Namun, data yang dipindahkan ini volumenya lebih kecil.



Data integration biasanya akan dimulai dengan proses penyerapan yang mencakup langkah-langkah seperti pembersihan data, pemetaan ETL (Extract, Transform, and Load), dan transformasi.

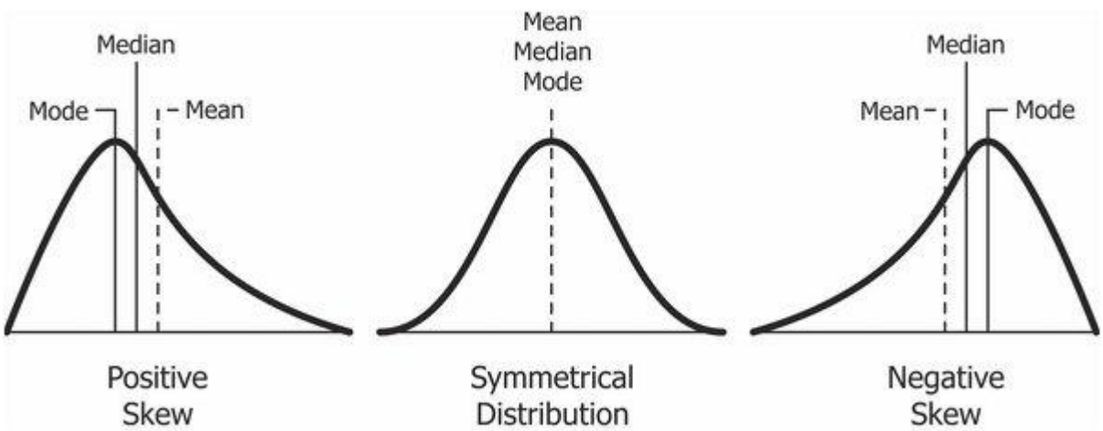
Data Transformation

Sekarang kita akan membahas tentang Data Transformasi. beberapa teknik Data Transformasi seperti mengubah tipe data, menghapus baris, memodifikasi kolom, menambah kolom baru, menghapus kolom dan mengurangi skew.

Dari masing-masing teknik transformasi, mengubah tipe data bisa menggunakan fungsi .astype("tipedata").

Lalu ada menghapus baris dengan cara menggunakan fungsi .drop contoh penggunaan : hapus baris index 0 = nama_table.drop(index = 0) apabila ingin menghapus banyak index sekaligus bisa memanfaatkan fungsi dari array.

Dan poin yang paling penting adalah penanganan skew seperti gambar dibawah ini :



Untuk treatmentnya sendiri apabila skew salah satunya bisa mengubah/transform data menggunakan fungsi median, IQR dll.

Normalization salah satu teknik data transformation diatas yaitu normalization, dengan cara mengurangi skala batas atas dan bawah dari range data. Adapun untuk teknik normalization yang sering digunakan yaitu MinMax dan ZScore.

DataFrame Manipulation & Combination

Salah satu teknik yang sering digunakan dalam data preparation adalah dengan menambahkan data. Caranya yaitu dengan menggunakan fungsi append ke dalam dataframe. dengan catatan penambahannya harus menggunakan dictionary. Untuk teknik penggabungan bisa menggunakan Merge Data Frame, yaitu menggabungkan 2 data frame menjadi 1 data frame berdasarkan kolom yang sama.