**Pokok Bahasan II**

**Data Pre-processing with Python**

**Kode Pokok Bahasan**: TIK.RPL03.001.002.01

**Deskripsi Pokok Bahasan**:

Membahas bagaimana pengolahan data sebelum diproses menggunakan bahasa Python. Mempersiapkan sebuah data mentah, agar dapat sesuai dengan spesifikasi untuk tahap processing data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Elemen Kompetensi | Indikator Kinerja | Jml Jam | Hal |
| 1 | Memahami data preprocessing menggunakan Python. | 1.1 Menganalisa data preprocessing dari data yang disediakan | 1 |  |
| 2 | Mengimplementasikan preprocessing data pada console Jupyter Notebooks. | 1.1 Mampu memproses data dengan Missing Values  1.2 Mampu menangani missing Value | 2 |  |

**TUGAS PENDAHULUAN**

Hal yang harus dilakukan dan acuan yang harus dibaca sebelum praktikum :

1. Menginstal Anaconda Python pada PC masing-masing praktikan.
2. Menginstal Jupyter Notebooks pada Anaconda Python untuk masing-masing praktikan.

**DAFTAR PERTANYAAN**

1. Apa yang dimaksud dengan Missing Value?
2. Adakah perbedaan penanganan Missing Value antara R dan Python, berikan Alasannya?

**JAWABAN**

**TEORI SINGKAT**

Data Preparation atau bisa disebut juga dengan data preprocessing adalah suatu proses/langkah yang dilakukan untuk membuat data mentah menjadi data yang berkualitas(input yang baik untuk data mining tools). Tahapan dalam melakukan data mining salah satunya adalah preprosesing data. Pertanyaannya adalah mengapa data perlu di bersihkan sebelum diproses? Hal ini terjadi karena biasanya data yang akan digunakan belum baik, penyebabnya antara lain :

* **Incomplete** : kekurangan nilai-nilai atribut atau atribut tertentu lainnya. – **Noisy** : berisi kesalahan atau nilai-nilai outlier yang menyimpang dari yang diharapkan.
* **Inconsisten** : ketidakcocokan dalam penggunaan kode atau nama.

Disini kualitas data yang baik didasarkan oleh keputusan yang baik dan data warehouse memerlukan integrasi kualitas data yang konsisten. Beberapa hal yang perlu diperhatikan untuk mendapatkan data yang baik adalah

:

–**Accuracy**

–**Completeness**

–C**onsistency** –**Timeliness**

* **Value added**

–**Interpretability**

–**Accessibility**

–**Contextual**

* **Representational**

**LAB SETUP**

Hal yang harus disiapkan dan dilakukan oleh praktikan untuk menjalankan praktikum modul ini.

1. Menginstall library yang dibutuhkan untuk mengerjakan modul.
2. Menjalankan R Studio.

**ELEMEN KOMPETENSI I**

**Deskripsi:**

Memahami data pre-processing menggunakan Python.

**Kompetensi Dasar**:

1. Menganalisa data pre-processing dari data yang telah disediakan.

**Latihan 1.1.1**

**Penjelasan Singkat :**

Pada latihan ini anda akan diminta untuk melakukan analisis terhadap data yang telah disediakan.

**Apakah data yang disediakan, termasuk data yang perlu di pre-processing? Jika ya, jelaskan alasannya!**

**ELEMEN KOMPETENSI** **2**

**Deskripsi:**

Mengimplementasikan pre-processing data.

**Kompetensi Dasar:**

1. Mampu memproses data dengan Missing Values
2. Mmapu menangani missing Value

**Latihan 1.2.1**

**Penjelasan Singkat:**

Pada latihan ini praktikan diminta untuk memproses missing value dari data yang telah disediakan. Bagaimana cara agar missing value tersebut tidak menjadi kendala dalam tahap processing menggunakan python.

**Langkah-Langkah Praktikum:**

1. Letakan dataset dalam sebuah folder, dataset adalah data train.csv yang didownload dari https://www.kaggle.com/c/titanic/data

2. Buka anaconda prompt dan jalankan perintah berikut ini :

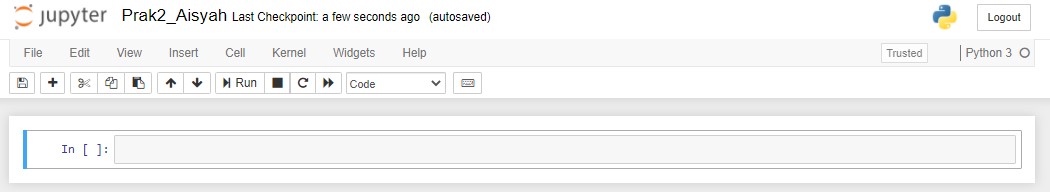
(base) C:\Users\Riva J>D:

(base) D:\>cd .semester7

(base) D:\.semester7>cd datmin

(base) D:\.semester7\datmin>jupyter notebook

1. Buat file baru pada jupyter notebook

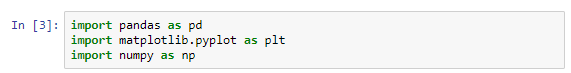


1. Import Library

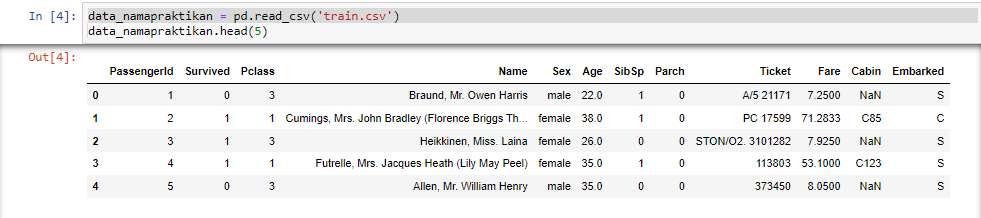
import pandas as pd

import matplotlib.pyplot as plt

import numpy as np



1. Mengambil data dari file .csv data\_namapraktikan=pd.read\_csv(‘train.csv’) data\_namapraktikan.head(5)



1. Adakah missing value pada variabe Age ? Y/T
2. Mengecek data secara umum dengan sintaks data\_namapraktikan.info()

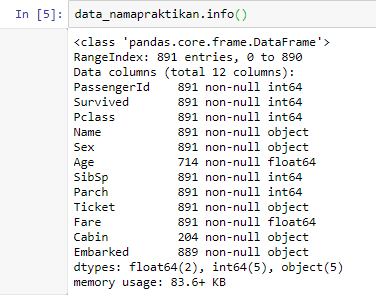
-ada berapa baris dan kolom?

-apa fitur/kolom yang ada pada data?

-apa tipe data dari masing-masing kolom?

-fitur/kolom apa yang mengandung data null?

-kunjungi <https://www.kaggle.com/c/titanic> dan dapatkan informasi lengkap dari setiap kolom yang ada.



1. Mengambil nama kolom dan tipe data dari masing-masing kolom

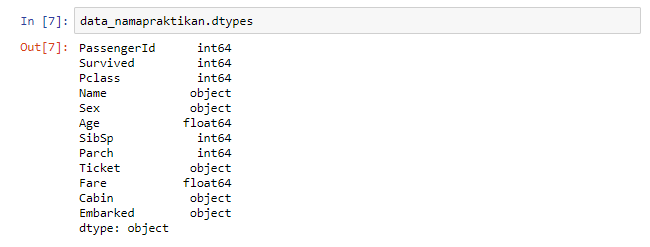
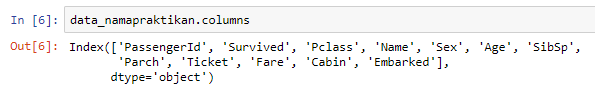
Selain dengan data\_frame.info(), pengambilan informasi kolom dapat dilakukan dengan

#data\_frame.columns

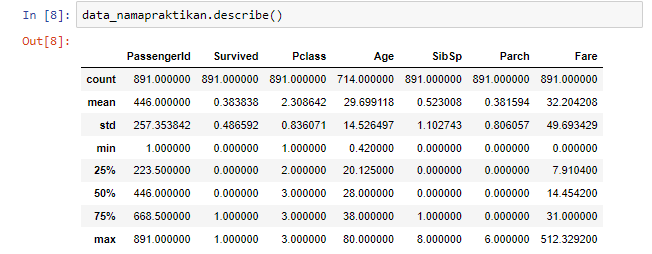
#data\_frame.dtypes

Dimana data\_frame adalah nama variabel yang menyimpan data dengan tipe data DataFrame

Cek apa representasi dari data bertipe object pada kolom/fitur Name, Sex, dan Ticket?

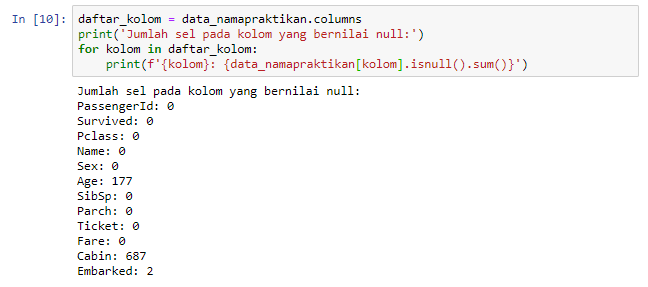


1. Mengidentifikasi nilai mean, median, min value, max value quartile dan standar deviasi pada data bertipe numerik.



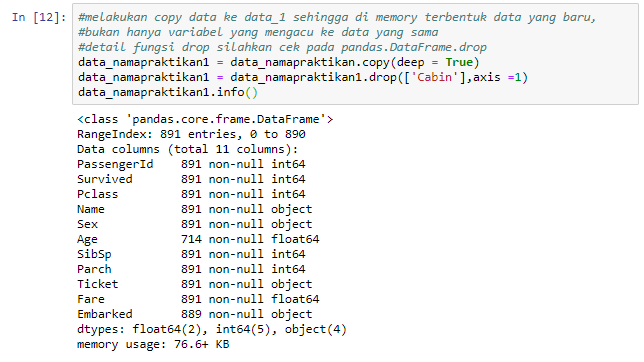
1. Cara lain mengecek missing value selain dengan data\_frame.info() mengambil nama kolom dan menghitung jumlah data pada kolom yang bernilai null dengan data[nama\_kolom].isnull().sum()

Kolom mana yang mengandung banyak nilai null?

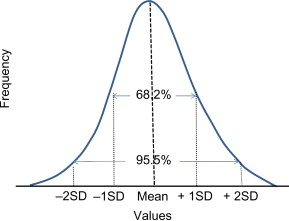


1. Data kolom Cabin mengandung banyak sekali nilai null, kita bisa mengabaikan kolom tersebut dengan cara tidak menyertakan data kolom Cabin dengan cara membuangnya. copy data kedalam variabel baru data\_1 dan panggil fungsi drop kolom

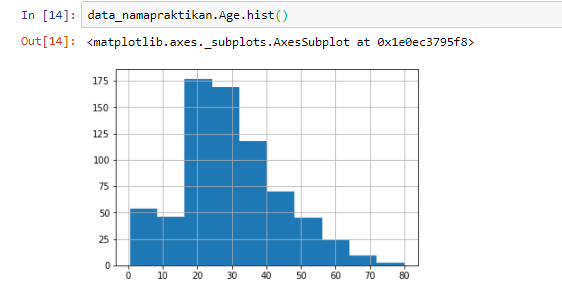
Pada fungsi drop parameter axis di set = 1. Apa maksudnya?



Data Data yang digunakan untuk membangun model seharusnya berdistribusi normal. Cara mudah untuk mengidentifikasi distribusi data adalah dengan membandingkan kurva distribusi dengan kurva gaussian yang berbentuk lonceng seperti gambar dibawah ini. Jika bentuknya tidak mirip dengan kurva berbentuk lonceng maka distribusi datanya tidak normal. (SD : standard deviation).



Cara memvisualisasikan distribusi data dengan mudah dilakukan dengan fungsi histogram.



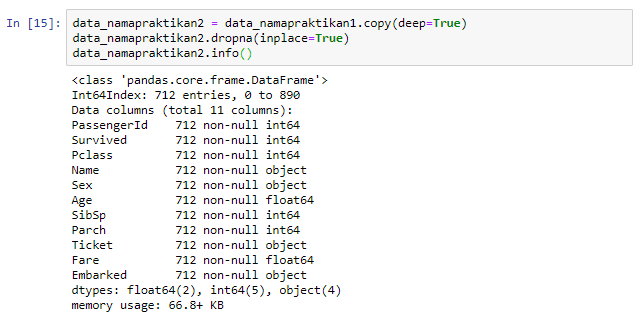
**Latihan 1.2.2**

Penjelasan Singkat:

Pada latihan ini praktikan diminta untuk mampu menangani missing value dengan data baru

Lakukan penanganan missing value pada variabel age dan salary sesuai langkahlangkah yang telah dipelajari dari data berikut :

1. Cara paling mudah untuk membuang nilai missing value adalah dengan memanggil fungsi dropna()



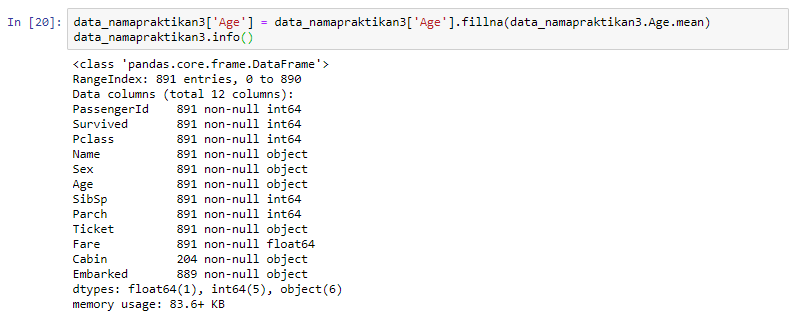
Bandingkan dengan data sebelum dilakukan pemanggilan fungsi dropna().

Apa yang terjadi setelah pemanggilan fungsi dropna()?

1. Melakukan pemberian nilai pada data yang missing value (imputasi)

Ada beberapa cara yang dapat dilakukan untuk melakukan pemberian nilai missing value:

* Memberikan nilai konstan
* Nilai yang diambil secara acak dari kolom yang sama
* Mean, median atau mode dari kolom tersebut
* Diestimasi dari model prediksi lain

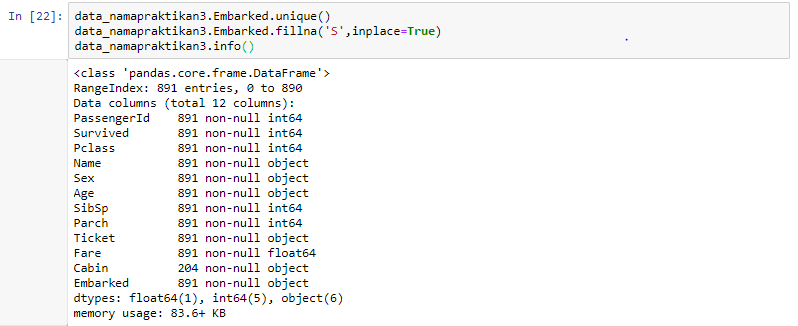


Perhatikan pada kolom Age, apakah masih ada data na? Perhatikan apa yang berubah dari data pada kolom Age?

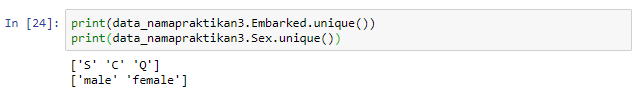
Lakukan pengisian pada nilai null data Embark dengan mengambil secara random dari baris data lain secara acak !

1. Mengubah representasi data kolom Sex dan Embark.

Lakukan pengisian data missing value pada kolom Embark dengan salah satu nilai selain ‘nan’ terlebih dahulu.

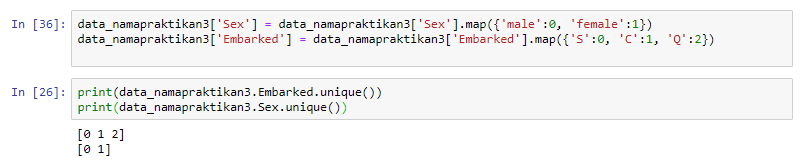


Cek nilai unik pada masing-masing kolom, kemudian petakan data pada kolom tersebut menjadi angka.

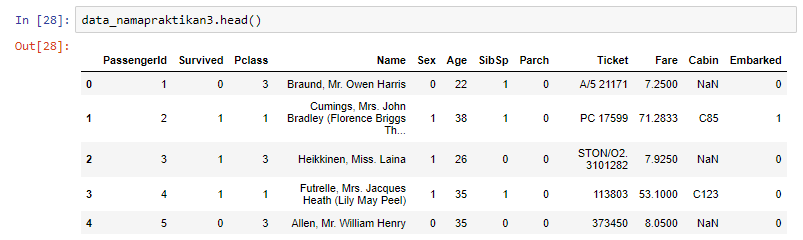


Ubah representasi dari (‘S’ ⇒ 0, ‘C’ ⇒ 1, ‘Q’ ⇒ 2)

Ubah representasi male ⇒ 0 dan female ⇒ 1



1. Setelah dilakukan beberapa preprocessing, tampilkan kembali beberapa baris data untuk mengecek ulang representasi data hasil proses cleaning



1. Nama dan Tiket bisa diabaikan dengan melakukan drop kolom. Lakukan drop kolom pada kedua kolom tersebut.

**Latihan 1.2.3**

Lakukan penanganan missing value pada variabel age dan salary sesuai langkahlangkah yang telah dipelajari dari data berikut :



Output

**CEK LIST**

Elemen

Kompetensi

No Latihan

Penyelesaian

S

elesai

Tidak selesai

1

1

.1.1

2

1

.2.1

1.2.2

**FORM UMPAN BALIK**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Elemen Kompetensi** | **Tingkat Kesulitan** | | | **Tingkat Ketertarikan** | | | **Waktu**  **Penyelesaian dalam menit** |
| Memahami data preprocessing menggunakan Python. |  | | |  | | |  |
|  |  | Sangat Mudah    Mudah    Biasa    Sulit    Sangat Sulit |  |  | Tidak Tertarik    Cukup Tertarik    Tertarik    Sangat Tertarik |
| √ |  |
|  |
|  |
|  |  |
|  | √ |
|  |
|  |
|  |
|  |
| Mengimplementasikan pre-processing data pada console Jupyter  Notebooks |  | | |  | | |  |
|  |  | Sangat Mudah    Mudah    Biasa    Sulit    Sangat Sulit |  |  | Tidak Tertarik    Cukup Tertarik    Tertarik    Sangat Tertarik |
| √ |  |
|  |
|  |
|  |  |
|  | √ |
|  |
|  |
|  |