DataFrame Basics and Data Preprocessing

Mhd. Arsya Fikri | 191402066

Apa yang akan kita bahas?

DataFrame Basics

Data Preprocessing

.

.

.

.

01

DataFrame Basics

Pandas

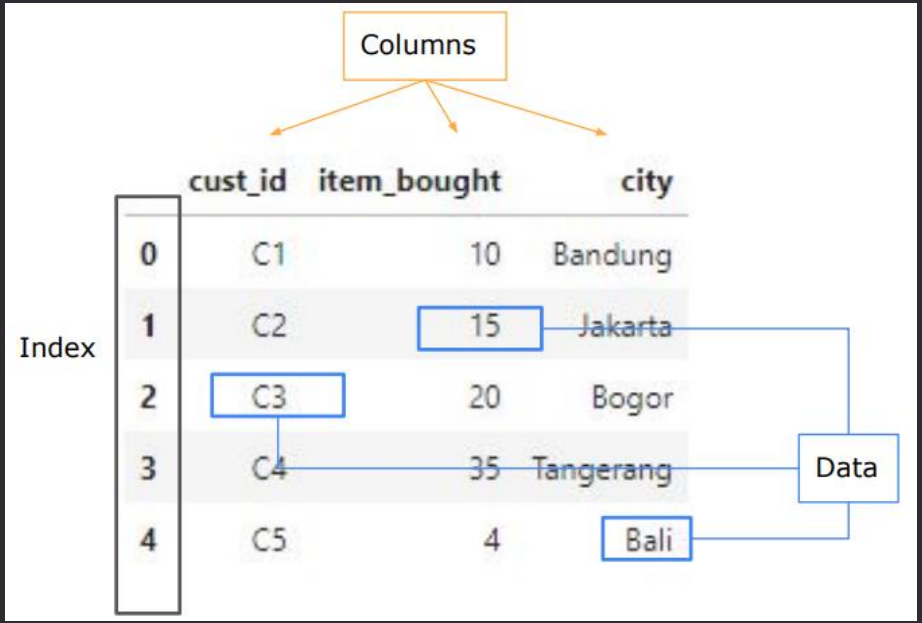
Pandas

- An open-sourced Python package that provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive.
- https://github.com/pandas-dev/pandas

Pandas

- Memiliki fitur-fitur yang luas
- Mempersingkat proses pengambilan data
- Dapat mengolah data dalam jumlah besar
- Menyediakan fitur representasi data
- Memiliki fitur penyesuaian data secara fleksibel
- Dapat dijalankan diberbagai *text editor*

Pandas DataFrame



DataFrame Basics

Hands On P2 – DataFrame Basics

02

Data Preprocessing

Data Preprocessing

Data Formatting

Handling Missing Values

Data Normalization

Data Binning

.

Data Formatting

- *Data formatting* merupakan proses membentuk data menjadi bentuk standar ekspresi yang umum agar lebih mudah dipahami. Ini merupakan salah satu proses *data cleaning* dalam *project data science*.

- Hands On P2 – Data Preprocessing

Handling Missing Values

- Raw data is not ideal. Many problems can create missing values in our dataset. As Data Scientists, missing values are common problems that we face, especially in the beginning stages of the project.
- What should we do?
- Ideally, the first thing we should do is: Ask the stakeholder that gives us the data.
- What if there's no one to ask? Before exploring our options, check if the missing values are intentional. For example, if a product has 0 sales on a particular date, then it makes sense that we might have missing value on that date.
- Hands On P2 – Data Preprocessing

Data Normalization

- Normalisasi data adalah proses membuat beberapa variabel memiliki rentang nilai yang sama, tidak ada yang terlalu besar maupun terlalu kecil sehingga dapat membuat analisis statistik menjadi lebih mudah. Perhatikan dua tabel berikut.

Tanpa Normalisasi

Umur	Gaji
20	100000
30	20000
40	500000
...	...

Dengan Normalisasi

Umur	Gaji
0.2	0.2
0.3	0.04
0.4	1
...	...

Data Normalization

- Ketika kita melakukan analisis lebih jauh seperti misalnya pemodelan menggunakan *linear regression*, variabel gaji akan lebih mempengaruhi hasil dikarenakan nilainya yang lebih besar. Model regresi linear akan menimbang gaji dengan lebih berat dari pada umur.

- Beberapa metode untuk normalisasi data di antaranya:
  - Standard Scaler
  - Min-Max Scaler
  - Robust Scaler
  - Normalizer
  - dll.

Data Normalization

- Salah satu metode normalisasi data adalah Min-Max Scaler. Cara kerjanya setiap nilai pada sebuah fitur dikurangi dengan nilai minimum fitur tersebut, kemudian dibagi dengan rentang nilai atau nilai maksimum dikurangi nilai minimum dari fitur tersebut. Cara ini akan menghasilkan nilai baru hasil normalisasi antara 0 sampai 1.

$$x_{new} = \frac{x_{old} - x_{min}}{x_{max} - x_{min}}$$

Data Normalization

- Metode selanjutnya adalah *Standard Scaler* atau disebut juga *Z-score*. Dengan formula ini, masing-masing nilai pada fitur dikurangi dengan miu (μ) yang merupakan nilai rata-rata fitur, kemudian dibagi dengan sigma (σ) yang merupakan standar deviasi. Cara ini akan menghasilkan nilai baru hasil normalisasi yang berkisar di angka 0 dan biasanya ada pada rentang antara -3 dan 3 tetapi bisa juga lebih tinggi atau lebih rendah.

$$x_{new} = \frac{x_{old} - \mu}{\sigma}$$

- Hands On P2 – Data Preprocessing

Data Binning

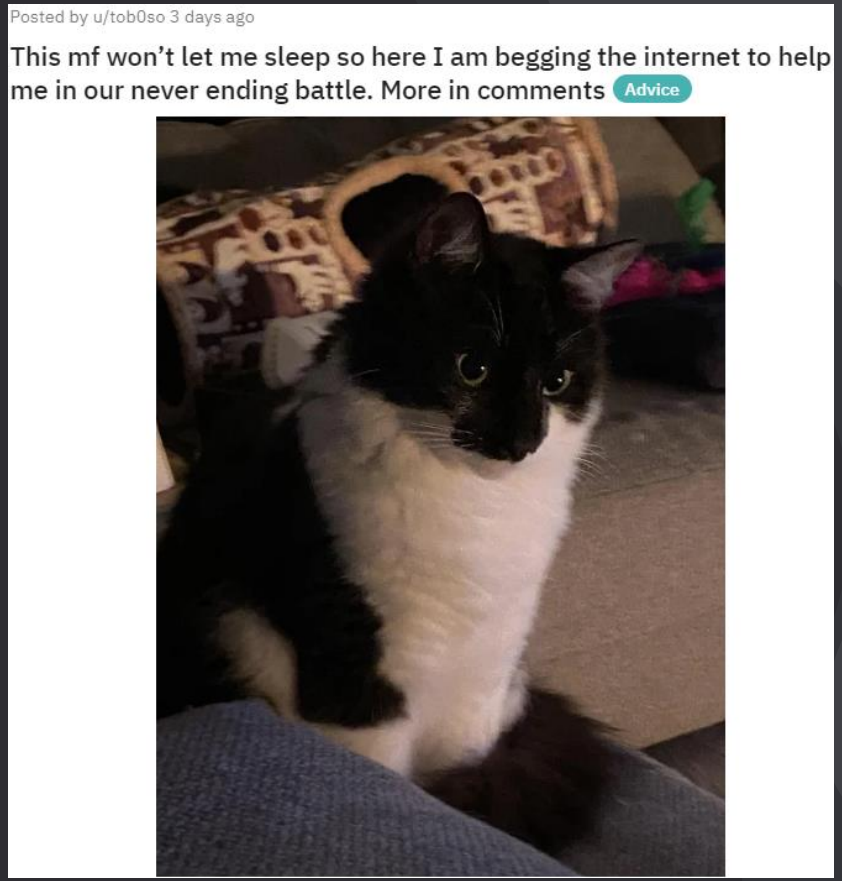
- Binning is a way to group a number of more or less continuous values into a smaller number of "bins". For example, if you have data about a group of people, you might want to arrange their ages into a smaller number of age intervals.
- Binning data merupakan salah satu teknik praproses data yang digunakan untuk meminimalisasi kesalahan dalam pengamatan serta terkadang dapat meningkatkan akurasi dari model prediktif.
- Hands On P2 – Data Preprocessing

Tugas :D

- Carilah sebuah dataset *supervised learning* yang mirip seperti contoh pada P2, namun tidak boleh menggunakan dataset pada P2 (Telco Customer Churn, Car Price, dan Titanic)!
- Sumber dataset bebas dari manapun.
- Dataset yang digunakan boleh sama, namun tidak boleh lebih dari 5 orang.
- Lakukan proses analisis dan *data preprocessing* pada dataset tersebut!
- Tugas dikerjakan di Notebook dilengkapi dengan penjelasan (berupa Markdown dan komentar kodingan) seperti pada contoh Hands-On P2!
- Berikan judul menggunakan Heading pada *cell* pertama dengan format: “Tugas P2 - Data Preprocessing | Nama - NIM”.
- Deadline pada 10 April 2023 pukul 01.59 WIT melalui e-learning. Format nama file: “Nama_NIM_Tugas_P2.ipynb”

Thanks!

Do you have any questions?



CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**