DATA PREPARATION: HOTEL BOOKING DEMAND

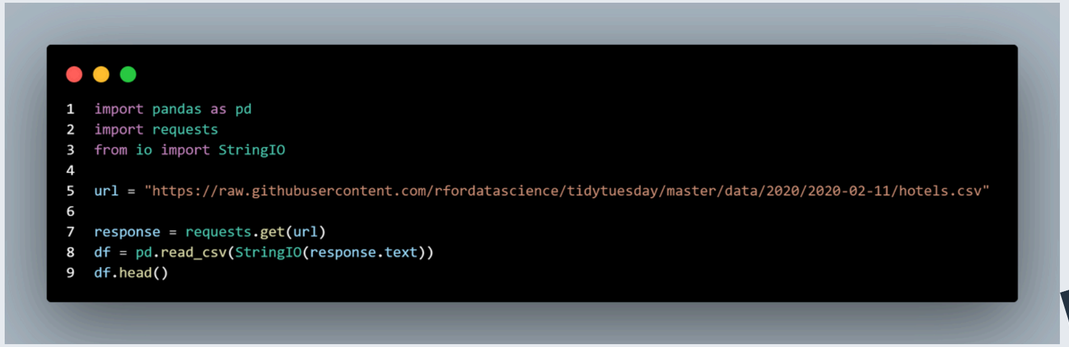
Oleh: Kelompok 3

Data Description

Dataset “Hotel Booking Demand” memberi informasi mengenai pola pemesanan hotel dan faktor-faktor yang memengaruhi tingkat pembatalan maupun keberhasilan check-in tamu.
Dataset memiliki 119.390 entri dengan 32 fitur (kolom) yang mencakup informasi tentang jenis hotel, status pembatalan, lead time, tanggal kedatangan, lama menginap, jumlah tamu, serta berbagai faktor lain yang relevan dalam industri perhotelan.
Link Dataset:
https://www.kaggle.com/datasets/jessemostipak/hotel-booking-demand/data

Data Loading

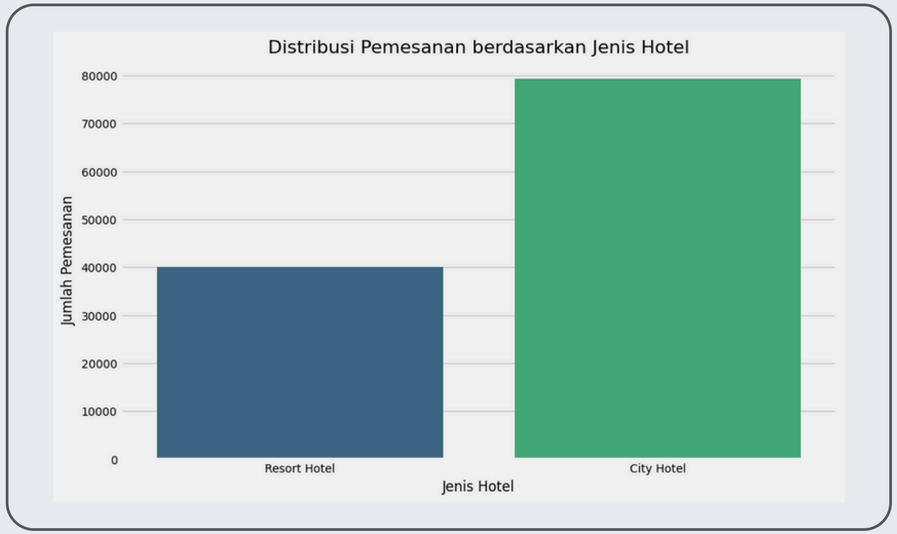
Pada bagian ini memuat dataset Hotel Booking Demand akan ke dalam lingkungan pemrograman Python. Proses ini menggunakan library Pandas untuk memuat dan memanipulasi data.

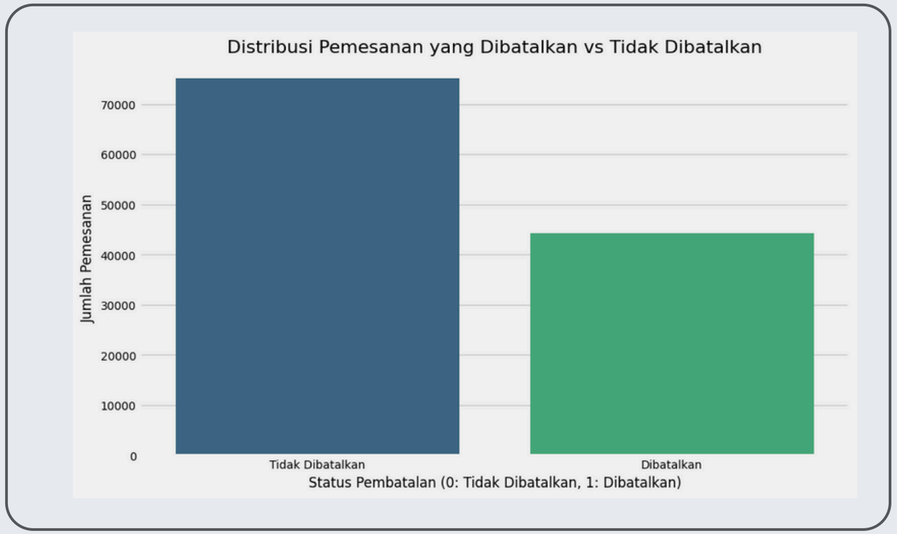


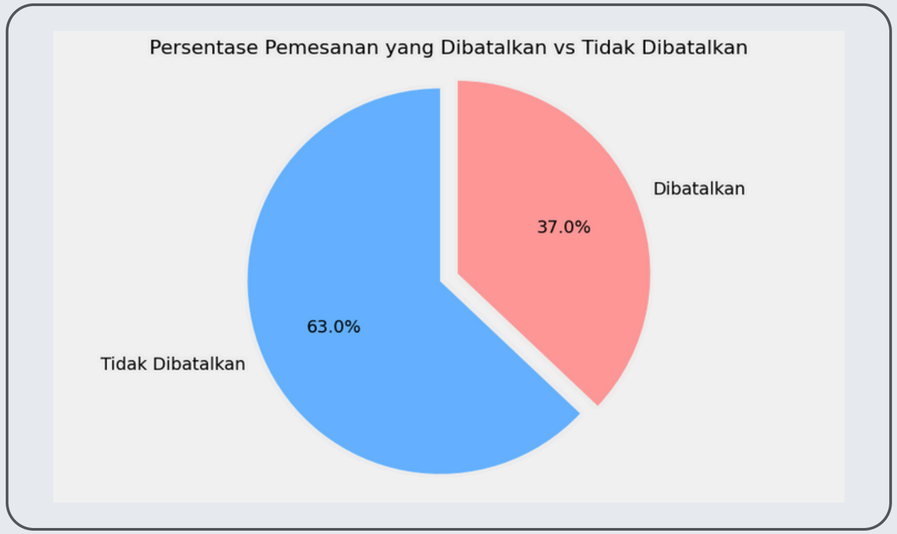
```python
import pandas as pd
import requests
from io import StringIO

url = "https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2020/2020-02-11/hotels.csv"

response = requests.get(url)
df = pd.read_csv(StringIO(response.text))
df.head()
```

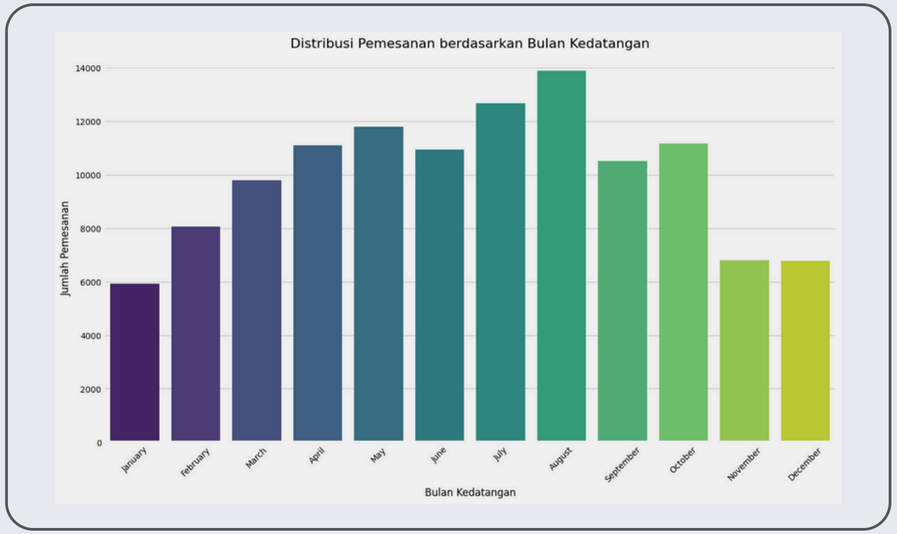
Data Understanding

Visualisasi Data



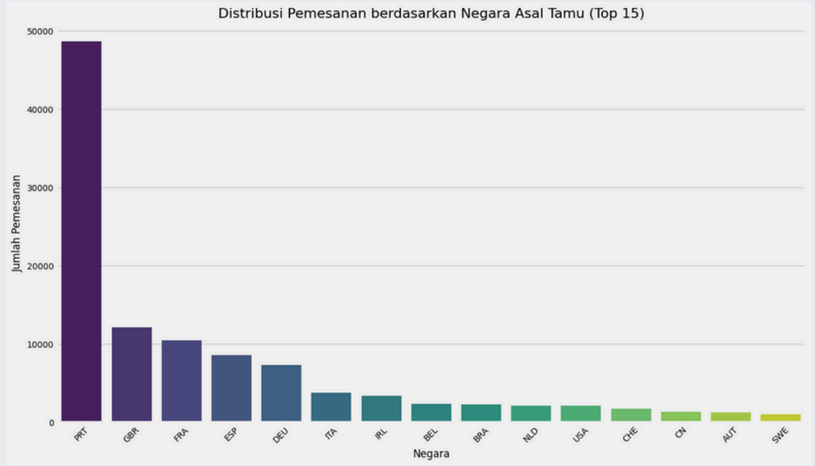


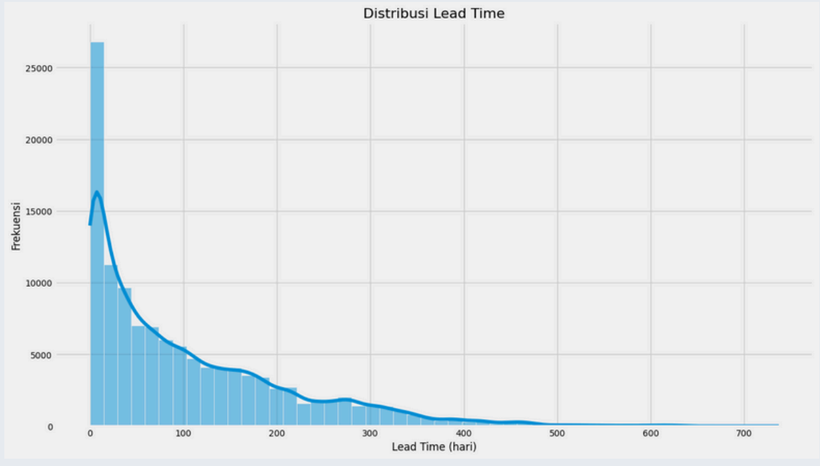


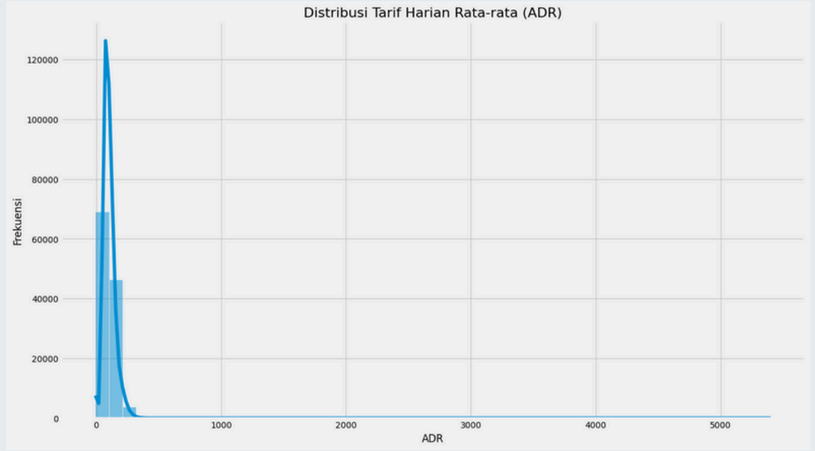


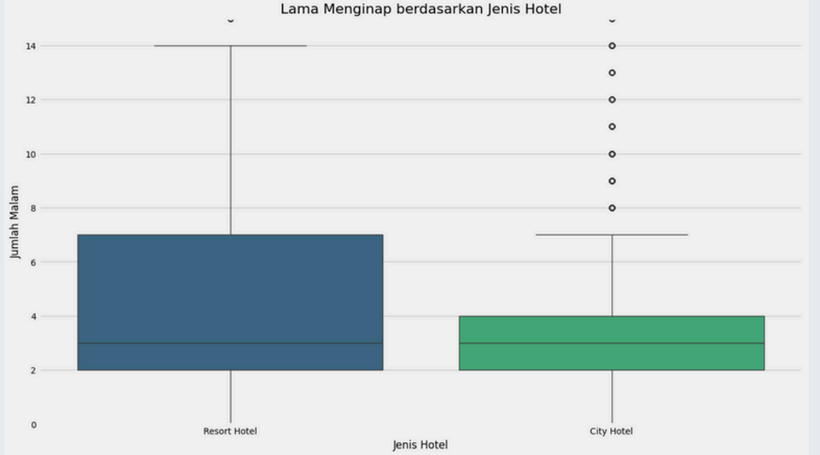
Data Understanding

Visualisasi Data









Data Preparation

Handling Missing Values

- Children: Diisi 0 jika kosong, asumsi tidak ada anak.
- Country: Diisi "Unknown" jika tidak diketahui.
- Agent & Company: Diisi 0, asumsi tidak melalui agen/perusahaan.

➡ Mempertahankan data tanpa menghapus informasi penting.

Encoding Categorical Variables

- Label encoding: Untuk kategori dengan urutan (arrival_date_month).
- Label encoding: Untuk kategori dengan banyak nilai (country).
- Mapping manual: Untuk kategori sedikit (hotel, meal, market_segment).

➡ Menyesuaikan metode encoding sesuai karakteristik data.

Handling Outliers

- ADR: Dibatasi 0 hingga persentil 99.5.
- Adults per room: Maksimal 5 orang.
- Lead time: Dibatasi hingga persentil 99.

➡ Mencegah pengaruh negatif outlier pada model.

Feature Engineering

- Total nights = stays_in_weekend_nights + stays_in_week_nights.
- Total guests = adults + children + babies.
- High season indicator, lead time category, weekend arrival flag.
- Special requests, repeated cancellations, price per person, room match.

➡ Menambah informasi yang dapat meningkatkan model.

Feature Selection

ANOVA F-value: Memilih 15 fitur paling relevan dengan is_canceled.

➡ Mengurangi dimensi data dan meningkatkan performa model.