DATA PREPARATION: HOTEL BOOKING DEMAND

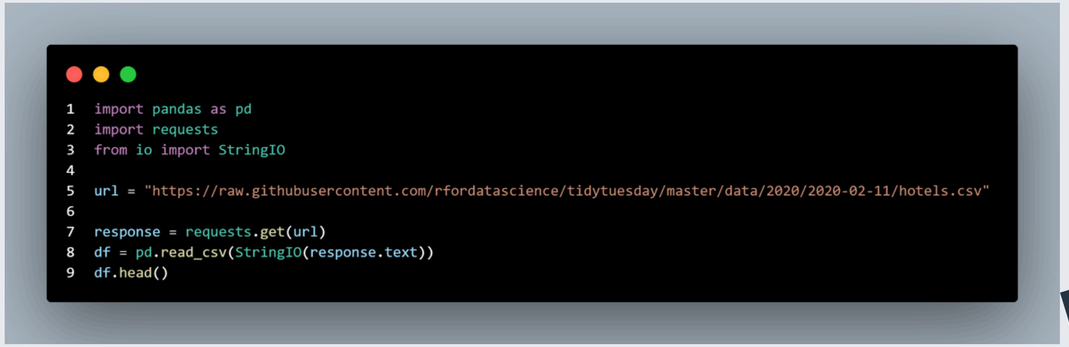
Oleh: Kelompok 3

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

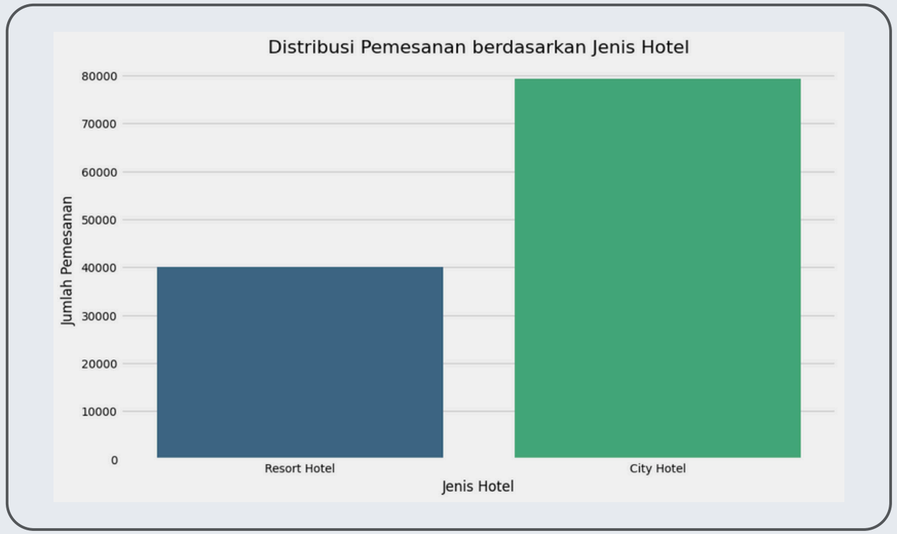
Pada bagian ini memuat dataset Hotel Booking Demand akan ke dalam lingkungan pemrograman Python. Proses ini menggunakan library Pandas untuk memuat dan memanipulasi data.

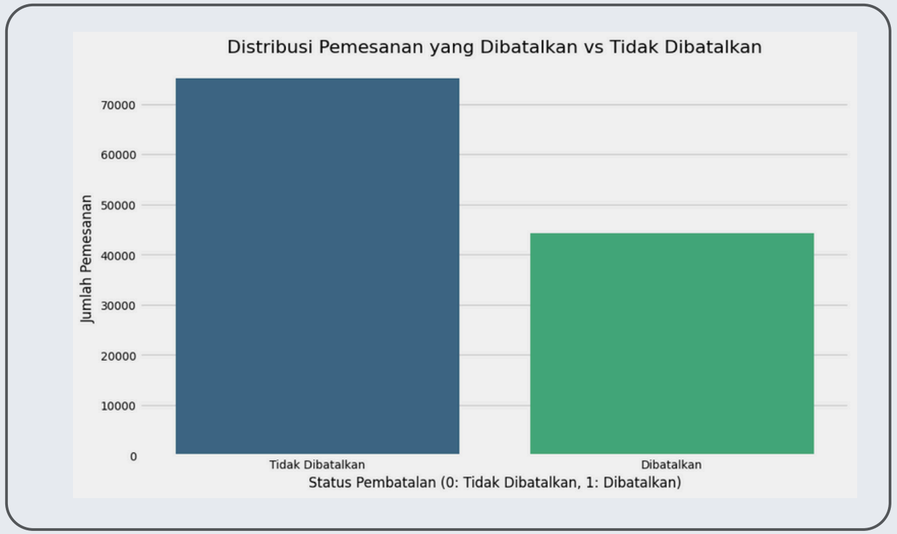


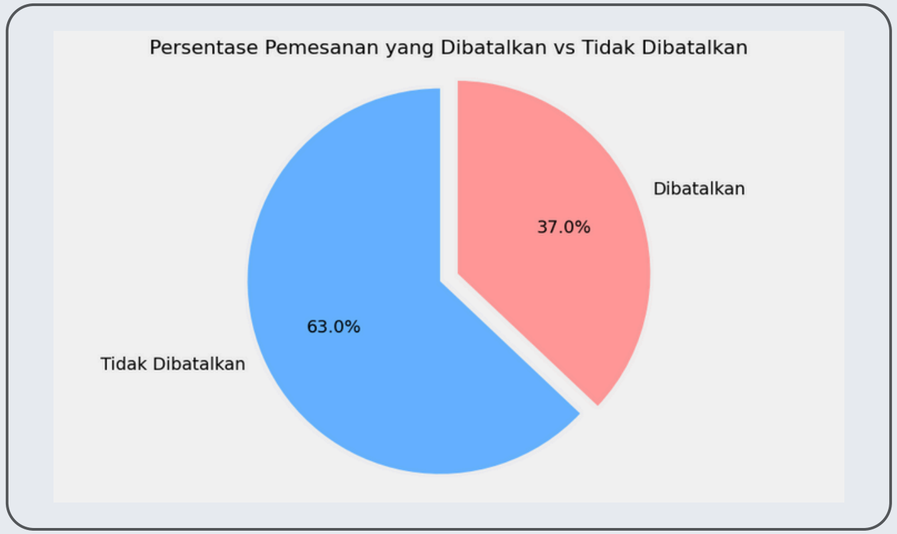
```python
import pandas as pd
import requests
from io import StringIO

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()
```

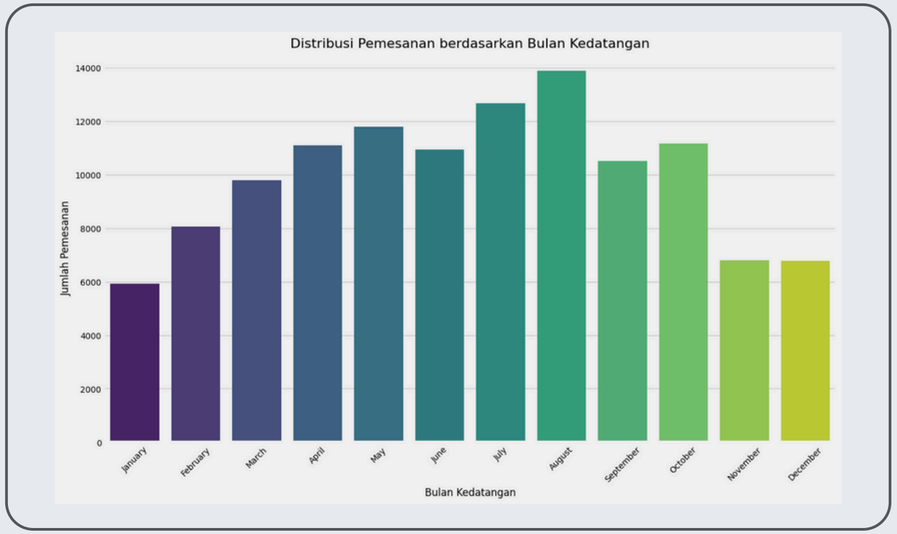
Data Understanding

Visualisasi Data



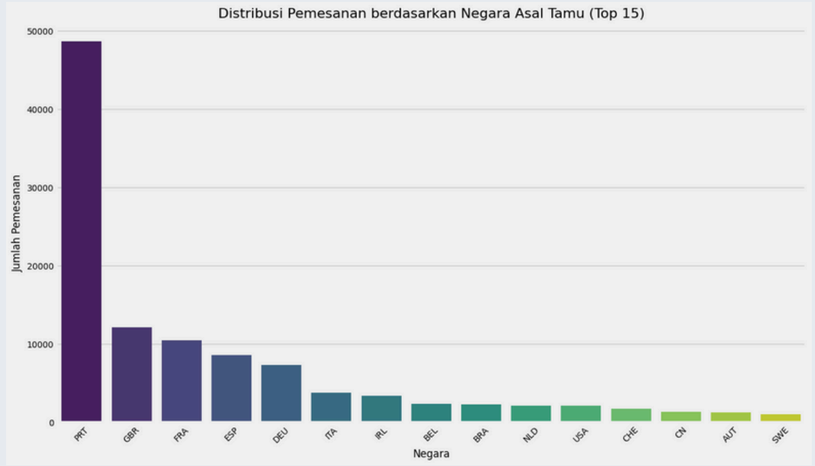


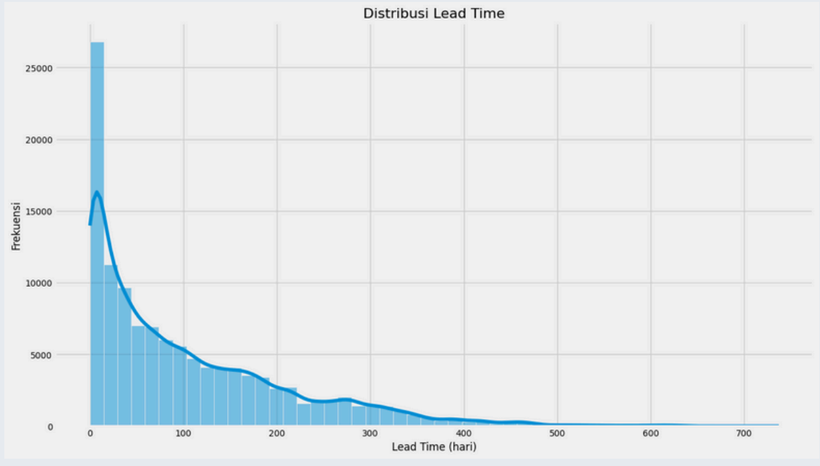


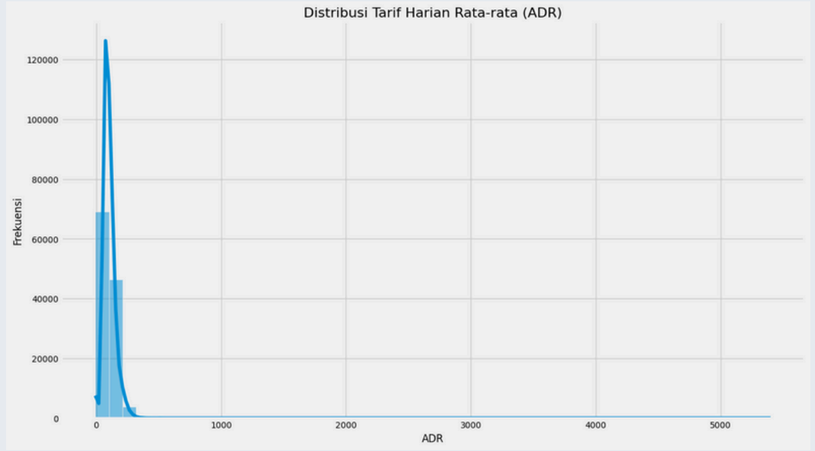


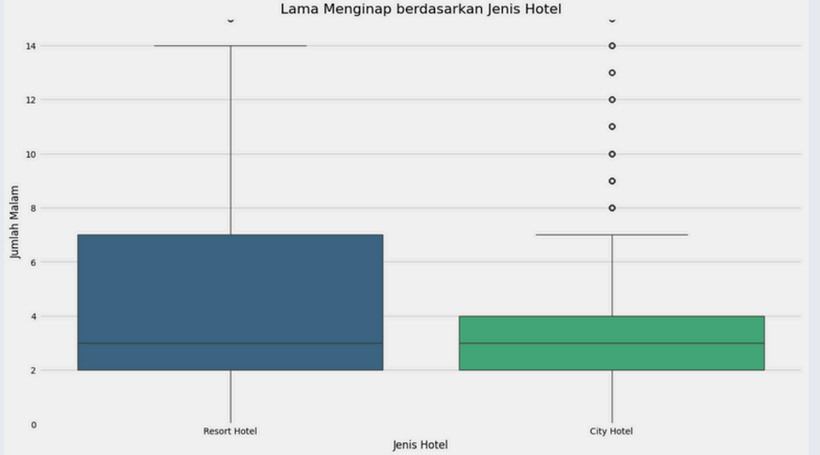
Data Understanding

Visualisasi Data









Data Preparation

Missing Values

mengisi nilai yang hilang dalam dataset dengan 0

Encoding

pengkodean variabel kategorikal menjadi numerik

Outliers

membatasi nilai maksimum ADR menggunakan persentil ke-99.5

Feature Engineering

melakukan rekayasa fitur untuk menciptakan fitur-fitur baru

Feature Selection

Menggunakan ANOVA F-value untuk memilih 15 fitur teratas yang paling relevan dengan variabel target 'is_canceled'