

Investigate Business Hotel using Data Visualization



Created by:
Achmad Hilman Shadiqin
ahilmans92@gmail.com
[linkedin](#)

Has more than 5 years of experience as an IT Support with special expertise in Network Administration. Decided to take a bold step in developing a career in data, focusing on analyzing and managing information to make a significant contribution to decision making and operational efficiency. The combination of technical expertise and motivation to continuously improve makes me ready to face new challenges.

“Sangat penting bagi suatu perusahaan untuk selalu menganalisa performa bisnisnya. Pada kesempatan kali ini, kita akan lebih mendalami bisnis dalam bidang perhotelan. Fokus yang kita tuju adalah untuk mengetahui bagaimana perilaku pelanggan kita dalam melakukan pemesanan hotel, dan hubungannya terhadap tingkat pembatalan pemesanan hotel. Hasil dari insight yang kita temukan akan kita sajikan dalam bentuk data visualisasi agar lebih mudah dipahami dan bersifat lebih persuasif.”

- Melakukan pengecekan informasi dari dataset, seperti jumlah baris dan kolom, tipe data, unique values setiap kolom, dan jumlah unique values setiap kolom.
- Melakukan pengecekan missing values, duplicates, dan invalid values
 - Menghapus kolom yang mempunyai terlalu banyak missing values (>50%).
- Feature Engineering pada kolom yang termasuk datetime
 - Year, Month, Date menjadi 1 kolom (arrival date) dan convert ke datetime object
 - Ditemukan date yang out of range, seperti 29 Februari dan 31 September
 - Menghapus baris yang datanya out of range, Februari (993 baris) dan September (365 baris)
 - Mengupdate kolom week of year dengan nilai baru dari arrival date
- Melakukan statistical analysis seperti descriptive statistics.
- Visualisasi distribusi data bertipe data numerik dan membuat perbandingan berdasarkan cancel dan jenis hotel.

Summary:

Berdasarkan pada proses Data Preprocessing, terdapat adanya inkonsistensi data seperti out of range dari data yang bertipe date (tanggal), adanya missing values di beberapa kolom dan terdapat yang lebih dari 50% pada 1 kolom, terdapat lebih dari 10% data yang terduplikasi, dan distribusi data numerik mayoritas terdistribusi positif.

Dimensi dataset setelah melalui proses data preprocessing dari 119390 baris, 29 kolom menjadi 118031 baris, 29 kolom atau turun sebesar 1.14%.