Homework

Exploratory Data Analysis

1. **Descriptive Statistics**

   Untuk mendapatkan ringkasan singkat tentang dataset, kami menggunakan fungsi info(). Hasil observasi yang didapatkan adalah sebagai berikut.

   A. Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?
      - Dari total 18 kolom, ada 4 kolom atau fitur dengan tipe data yang kurang sesuai, yaitu `OperatingSystems`, `Browser`, `Region`, dan `TrafficType`, seharusnya *string* bukan *integer*, karena kemungkinan sudah melalui proses *label encoding*, sedangkan kolom lainnya sudah sesuai.

   B. Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?
      - Tidak ada kolom yang memiliki nilai kosong atau *missing values*.

   C. Apakah ada kolom yang memiliki nilai summary agak aneh?
      - Dalam fitur numerikal, ada perbedaan yang signifikan antara nilai mean dan median (P50), yaitu nilai mean > median, karena kemungkinan dipengaruhi oleh adanya ***outlier*** atau pencilan, sehingga distribusi data akan cenderung menceng ke kanan atau ***positively skewed***.
      - Dalam fitur kategorikal, 5 dari 8 fitur memiliki kategori yang sangat mendominasi (dengan proporsi > 50%).

2. **Univariate Analysis**

   Berdasarkan *univariate analysis* yang dilakukan, didapatkan hasil observasi sebagai berikut.

   - Sebagian besar fitur memiliki distribusi yang ***positively skewed***, karena nilai mean > median.
   - Sebagian besar fitur memiliki ***outlier*** atau pencilan.
   - Fitur `OperatingSystem` distribusinya **multimodal**, karena nilai mode > 2.
   - Fitur `Month` distribusinya mendekati **bimodal** dengan data tertinggi pada bulan Mei dan November.

- Fitur `VisitorType` dengan nilai Returning_Visitor sangat mendominasi.
- Fitur `Weekend` dengan nilai False mendominasi.
- Fitur `Revenue` dengan nilai False (tidak melakukan *purchasing*) sangat mendominasi.
- Fitur `Browser` dan `TrafficType` memiliki kategori yang cukup banyak (> 10 kategori).

Oleh karena itu, pada saat data pre-processing, kita perlu:

- Menghapus *outlier* pada setiap fitur bisa menggunakan IQR atau Z-Score.
- Melakukan *Data Transformation* dengan *Log Transformation* atau menggunakan teknik lain yang paling sesuai, karena terdapat banyak fitur yang memiliki sebaran *right skew*.
- Melakukan *Feature Encoding* untuk fitur `Month`, `Weekend`, dan `Revenue` menggunakan *Label Encoding*, sedangkan untuk fitur `VisitorType` menggunakan One Hot Encoding, karena terdapat nilai > 2 dan bukan tipe ordinal.
- Melakukan *Handling Imbalanced Data* untuk fitur `Revenue`, karena fitur ini merupakan target yang mempunyai ketimpangan data yang signifikan.

3. **Multivariate Analysis**

Berdasarkan *multivariate analysis* yang dilakukan, didapatkan hasil observasi sebagai berikut.

A. Bagaimana korelasi antara masing-masing fitur dan label. Kira-kira fitur mana saja yang paling relevan dan harus dipertahankan?
    - Banyak fitur tidak berkorelasi.
    - Fitur `PageValues` dan `Revenue` cukup berkorelasi (0,49) dan memiliki pola hubungan ***positive linear association***, karena ketika tidak ada pembelian yang dilakukan, jumlah sesi dengan Page Values = 0 relatif tinggi, yaitu sebanyak 9.230 sesi.

- Tingginya nilai `PageValues` berbanding lurus dengan naiknya nilai `Revenue`, sehingga fitur `PageValues` harus dipertahankan.

B. Bagaimana korelasi antar-fitur, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap fitur itu?
   - Fitur `BounceRates` dan `ExitRates` berkorelasi tinggi (0,91), karena ketika Bounce Rate meningkat, Exit Rate juga meningkat berdasarkan hasil perhitungan oleh Google Analytics, sehingga kita bisa memilih salah satu fitur, yaitu yang memiliki *correlation* lebih besar (`ExitRates`) atau bisa juga melakukan *Principal Component Analysis* (PCA).
   - Untuk menghasilkan `Revenue`, maka harus memiliki *Bounce Rates* yang rendah, *Exit Rates* yang rendah, dan *Page Values* yang tinggi.
   - Korelasi antara durasi atau waktu yang dihabiskan pelanggan di halaman tertentu terhadap halamannya terlihat cukup jelas.
   - Fitur `Administrative_Duration`, `Informational_Duration`, dan `ProductRelated_Duration` memiliki elemen unik atau kategori yang cukup banyak, sehingga kita bisa melakukan *feature selection/dimensionality reduction* atau kita bisa buat fitur baru `TotalPage_Duration` atau `AvgPage_Duration`.
   - Fitur `Administrative`, `Informational`, dan `ProductRelated` juga bisa buat fitur baru `TotalPage` atau `AvgPage`.

4. Business Insights

Tuliskan business insights, dan berdasarkan insight tersebut jelaskan rekomendasinya untuk bisnis.

A. Di daerah **`Region` 1** memiliki jumlah pengunjung situs web *e-commerce* yang terbanyak. Solusi untuk meningkatkan ketertarikan pengunjung, kita bisa melakukan promosi ke daerah-daerah yang jarang mengunjungi situs web dengan memberikan penawaran spesial, seperti gratis ongkos pengiriman (ongkir).

B. Pengunjung yang berkunjung pada ***weekend*** **lebih sedikit** dibandingkan dengan hari-hari biasa atau ***weekday***, sehingga kita bisa mengadakan *event* untuk menarik pelanggan melakukan transaksi pada waktu *weekend*.
C. Bagi ***Returning Visitor*** atau pelanggan yang sering berkunjung ke situs web dan melakukan transaksi akan mendapatkan kupon gratis belanja sebesar Rp50.000 yang dapat digunakan pada transaksi berikutnya.
D. Bagi ***New Visitor*** atau pelanggan baru, agar melakukan transaksi pertama, solusi kita, bisa diberikan produk gratis dengan syarat melakukan pembelanjaan sejumlah tertentu.
E. Pada fitur `Month`, diketahui bahwa bulan **Maret, Mei, November, dan Desember** merupakan bulan-bulan yang **sering dikunjungi** pengunjung, Solusi kita, coba untuk mengadakan suatu *event* di setiap bulan seperti event (yang dilakukan kompetitor) 1.1 hingga 12.12.
F. Pada bulan **Februari**, jumlah pelanggan yang mengunjungi situs web sangat sedikit dan terlihat dari revenue yang dihasilkan juga sedikit. Solusi kita, diberikan promo di hari Valentine untuk menarik minat pelanggan melakukan transaksi.

5. Git

a. Link **Repository** GitHub: https://github.com/sabirinID/Final-Project-Quattro/
b. File **Jupyter Notebook** yang di-upload: 1-Exploratory_Data_Analysis.ipynb