Stage 1

Exploratory Data Analysis

1. **Data Exploration**
   Dataset ***Online Shoppers Purchasing Intention*** merupakan dataset yang dibentuk secara khusus, sehingga setiap sesi akan dimiliki oleh pelanggan yang berbeda selama periode 1 tahun. Dataset ini terdiri dari 12.330 baris dan 18 kolom fitur, setiap baris berisi data yang berkaitan dengan sesi kunjungan (waktu yang dihabiskan) pelanggan pada situs e-commerce.

2. **Data Understanding**
   a. **Data Dimension**
      Dataset ini memiliki dimensi data, yaitu
      Jumlah baris: 12.330
      Jumlah kolom: 18

   b. **Data Types and Structure**
      Untuk mendapatkan ringkasan singkat tentang dataset, kami menggunakan fungsi info(). Hasil observasi yang didapatkan adalah sebagai berikut.
      - 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.
      - Tidak ada kolom yang memiliki nilai kosong atau missing values.
      - Tipe data berupa boolean (2), float (7), integer (7), dan string (2).

   c. **Detect Missing Values**
      Untuk memastikan adanya missing values dalam dataset, kita menggunakan metode isna().
      - Tidak ada kolom yang null (bernilai None ataupun NaN).

d. **Detect Duplicates**
   Untuk menemukan adanya duplicates, kita menggunakan metode duplicated(). Ternyata ditemukan data duplikat sebanyak 125 baris. Walaupun demikian, kita berasumsi bahwa data tersebut merupakan data unik, yang terkait dengan sesi kunjungan pelanggan.

e. **Unique Elements**
   Untuk mencari elemen unik dalam dataset, kita menggunakan fungsi nunique().
   - 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`.
   - Fitur `BounceRates`, `ExitRates`, dan `PageValues` akan dipertahankan.

3. **Descriptive Statistics**
   Untuk mendapatkan perincian statistik dasar dari dataset, kita menggunakan metode describe().
   a. **Numerical Features**
      Dalam fitur numerikal, ada perbedaan yang signifikan antara nilai mean dan median (P50), yaitu mean > median, karena kemungkinan dipengaruhi oleh adanya *outlier* atau pencilan, sehingga distribusi data akan cenderung menceng ke kanan atau *positively skewed*.

   b. **Categorical Features**
      Berikut ini nilai yang paling umum dalam fitur kategorikal, berturut-turut adalah:
      - `Month` : May (27,3%),
      - `OperatingSystems` : 2 (53,5%),
      - Browser` : 2 (64,6%),
      - `Region` : 1 (38,8%),

- `TrafficType` : 2 (31,7%),
- `VisitorType` : Returning_Visitor (85,6%),
- `Weekend` : 0 atau False (76,7%), dan
- `Revenue` : 0 atau False (84,5%).

Dalam fitur kategorikal, 5 dari 8 fitur memiliki kategori yang sangat mendominasi (dengan proporsi > 50%).

c. Target Feature

Fitur `Revenue` digunakan sebagai *target feature* atau label kelas.

- Dari total 12.330 sesi, 84,5% atau 10.422 sesi merupakan kelas negatif yang tidak diakhiri dengan pembelian, sedangkan 15,5% sisanya atau 1.908 sesi merupakan kelas positif yang diakhiri dengan pembelian.
- Dataset *imbalance* atau tidak seimbang, karena proporsi data minoritas (dalam hal ini kelas positif) relatif rendah, dengan *degree of imbalance:* ***moderate***.
- Pada saat ***data pre-processing***, kita perlu melakukan ***handling imbalance data***, seperti
  - Oversampling: menduplikasi data minoritas,
  - Undersampling: menghapus data mayoritas, atau
  - Class weight

4. Exploratory Data Analysis

a. 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.

b. Multivariate Analysis

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

- 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.
- 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.

5. Business Insights

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.

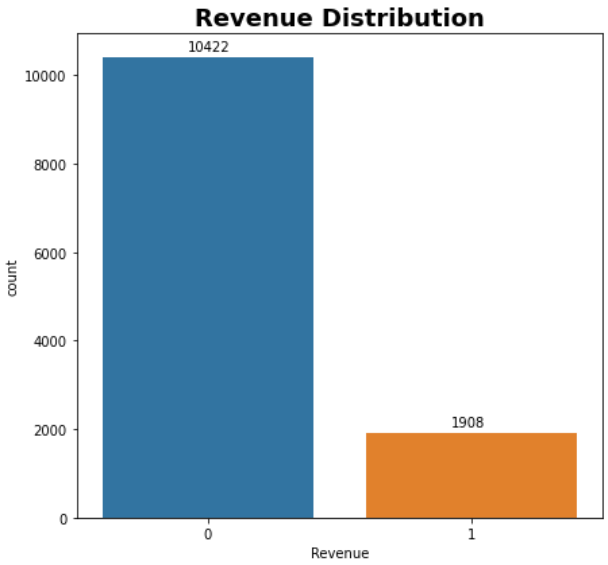
d. Bagi ***New Visitor*** atau pelanggan baru, agar melakukan transaksi pertama, solusi kita, bisa diberikan produk gratis dengan syarat melakukan pembelanjaan sejumlah tertentu.

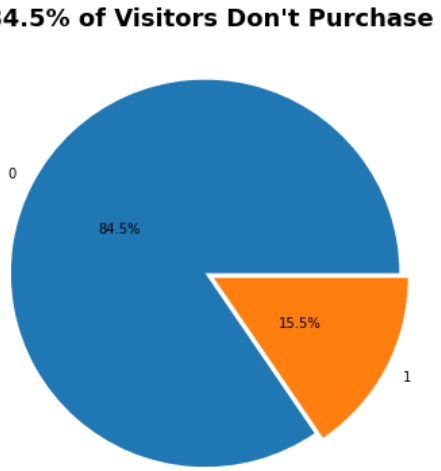
e. Pada fitur `Month`, diketahui bahwa bulan **Maret, Mei, November, dan Desember** merupakan bulan-bulan yang **sering dikunjungi** pengunjung, Solusi kita, coba untuk mengadakan suatu *event* di setiap bulan seperti event (yang dilakukan kompetitor) 1.1 hingga 12.12.

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.

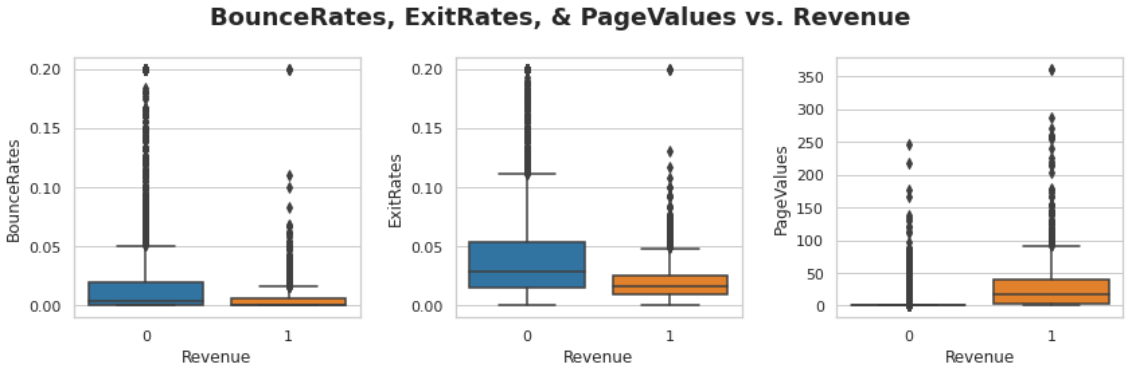
6. Data Visualization





Dari total 12.330 sesi, 84,5% atau 10.422 sesi merupakan **kelas negatif** yang tidak diakhiri dengan pembelian (Revenue = 0), sedangkan 15,5% sisanya atau 1.908 sesi merupakan **kelas positif** yang diakhiri dengan pembelian (Revenue = 1).

Target *imbalance* atau tidak seimbang, karena proporsi data minoritas (dalam hal ini kelas positif) relatif rendah, dengan *degree of imbalance*: moderate.



Seperti yang ditunjukkan pada *boxplot* di atas, untuk menghasilkan Revenue, maka harus memiliki Bounce Rates yang rendah, Exit Rates yang rendah, dan Page Values yang tinggi.

7. **Data Storytelling**

   Dataset ***Online Shoppers Purchasing Intention*** berisi informasi tentang perilaku dan minat pembeli saat *browsing* situs web *e-commerce*. Dataset ini mencakup 18 kolom atau fitur, numerikal (10 fitur) dan kategorikal (8 fitur), termasuk jenis pengunjung, durasi sesi, dan jumlah halaman yang dilihat.

   Setelah menganalisis dataset, kami menemukan bahwa sebagian besar pengunjung merupakan *returning visitor*, dan hanya sedikit yang merupakan tipe pengunjung baru atau lainnya. Selain itu, pengunjung umumnya menjelajahi situs web di bulan Mei, November, dan Maret. Namun, data pada bulan Januari dan April tidak tersedia, kemungkinan terjadi kesalahan sistem, sehingga data sesi kunjungan pelanggan pada bulan-bulan tersebut tidak tercatat.

   Ketika menganalisis perilaku pengunjung, kami menemukan bahwa sebagian besar pengunjung hanya melihat beberapa halaman dan menghabiskan waktu relatif sedikit di situs web. Selain itu, kami menemukan bahwa *bounce rate*, yang mengukur persentase pengunjung yang meninggalkan situs web setelah hanya melihat satu halaman, relatif tinggi. Ada kemungkinan halaman web memiliki desain UI (*user interface*) dan UX (*user experience*) kurang menarik.

   Di sisi lain, kami menemukan bahwa pengunjung yang melihat lebih banyak halaman, menghabiskan lebih banyak waktu di situs web, dan memiliki nilai *page values* yang lebih tinggi, lebih cenderung melakukan pembelian. Selain itu, kami menemukan bahwa pengunjung yang *browsing* pada akhir pekan lebih cenderung melakukan pembelian daripada mereka yang *browsing* pada hari kerja.

Dengan menganalisis korelasi antara variabel, kami menemukan bahwa jumlah halaman yang dilihat dan durasi sesi sangat berkorelasi, mengindikasikan bahwa pengunjung yang melihat lebih banyak halaman cenderung menghabiskan lebih banyak waktu di situs web. Selain itu, kami menemukan korelasi negatif antara bounce rate dan page value, mengindikasikan bahwa pengunjung yang menemukan konten yang lebih berharga di situs web lebih sedikit kemungkinannya untuk pergi setelah melihat satu halaman.

Secara keseluruhan, hal ini menunjukkan bahwa meningkatkan fitur dan pengalaman pengguna dari situs web dapat mengarah pada peningkatan pendapatan dengan mengurangi *bounce rate* dan mendorong pengunjung untuk menjelajahi lebih banyak halaman dan menghabiskan lebih banyak waktu di situs web. Selain itu, menargetkan pengunjung pada weekday dan memberikan fitur yang relevan dan berharga juga dapat meningkatkan kemungkinan melakukan pembelian.