Rakamin Academy

Final Project (Stage 1)

Anggota Kelompok:
Muhammad Gilang Mahardika
Muhammad Farhan Al Hafizh
Suny Guinesya Ardiansyah
Aldi Pajar Romadon
Anuar Ali Syabana
Rivani Putra Irwanto
Agus Setiana

1. Descriptive Statistics

A. Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?

```
data_product.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12946 entries, 0 to 12945
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12835 non-null  float64
 1   Administrative_Duration  12313 non-null  float64
 2   Informational            12946 non-null  int64
 3   Informational_Duration   12946 non-null  float64
 4   ProductRelated           12946 non-null  int64
 5   ProductRelated_Duration  12307 non-null  float64
 6   BounceRates              12872 non-null  float64
 7   ExitRates                12946 non-null  float64
 8   PageValues               12946 non-null  float64
 9   SpecialDay               12946 non-null  float64
 10  Month                    12946 non-null  object
 11  OperatingSystems         12422 non-null  float64
 12  Browser                  12946 non-null  int64
 13  Region                   12946 non-null  int64
 14  TrafficType              12946 non-null  int64
 15  VisitorType              12946 non-null  object
 16  Weekend                  12946 non-null  bool
 17  Revenue                  12946 non-null  bool
dtypes: bool(2), float64(9), int64(5), object(2)
memory usage: 1.6+ MB
```

Seperti kita lihat di sini bahwa beberapa kolom memiliki tipe data yang kami rasa kurang sesuai, kolom -kolom tersebut adalah OperatingSystems, Browser, Region, TrafficType yang mana menurut kami kolom-kolom tersebut merupakan kolom bertipe kategori (object) tetapi pada dataframe ini masih bertipe numerik (int, float)

B. Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?

```
missing_values = data_product.isnull().sum()

columns_with_missing_values = missing_values[missing_values > 0]
print(columns_with_missing_values)

total_rows = len(data_product)
percentage_missing = (columns_with_missing_values / total_rows) * 100
percentage_missing = percentage_missing.round(2)  # membulatkan persenta
percentage_missing = percentage_missing.astype(str) + '%'  # menambahkan

print("\nPersentase data yang hilang per fitur:")
print(percentage_missing)
```

```
Administrative             111
Administrative_Duration    633
ProductRelated_Duration    639
BounceRates                 74
OperatingSystems           524
dtype: int64

Persentase data yang hilang per fitur:
Administrative             0.86%
Administrative_Duration    4.89%
ProductRelated_Duration    4.94%
BounceRates                0.57%
OperatingSystems           4.05%
dtype: object
```

Dalam dataset tersebut, terdapat missing value pada beberapa fitur dengan jumlah dan persentase yang berbeda:

1. **Administrative:**
Terdapat 111 missing value, yang mewakili sekitar 0.86% dari total data pada fitur tersebut.

2. **Administrative_Duration**:
Terdapat 633 missing value, yang mewakili sekitar 4.89% dari total data pada fitur tersebut.

3. **ProductRelated_Duration**:
Terdapat 639 missing value, yang mewakili sekitar 4.94% dari total data pada fitur tersebut.

4. **BounceRates**:
Terdapat 74 missing value, yang mewakili sekitar 0.57% dari total data pada fitur tersebut.

5. **OperatingSystems**:
Terdapat 524 missing value, yang mewakili sekitar 4.05% dari total data pada fitur tersebut.

C. Apakah ada kolom yang memiliki nilai summary agak aneh? (min/mean/median/max/unique/top/freq)

```
[ ] data_product.describe()
```

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	Operati
count	12835.000000	12313.000000	12946.000000	12946.000000	12946.000000	12307.000000	12872.000000	12946.000000	12946.000000	12946.000000	12946.000000	124
mean	2.303857	80.370267	0.498841	34.136048	31.657655	1192.740077	0.022309	0.043266	5.875963	0.061270	5.159200	
std	3.314427	175.494016	1.263276	140.022848	44.202635	1910.216261	0.048681	0.048808	18.414670	0.198667	2.371631	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.000000	0.000000	0.000000	0.000000	7.000000	182.083333	0.000000	0.014286	0.000000	0.000000	5.000000	
50%	1.000000	7.000000	0.000000	0.000000	18.000000	599.583333	0.003112	0.025329	0.000000	0.000000	6.000000	
75%	4.000000	92.933333	0.000000	0.000000	38.000000	1470.522917	0.016933	0.050000	0.000000	0.000000	7.000000	
max	27.000000	3398.750000	24.000000	2549.375000	705.000000	63973.522230	0.200000	0.200000	361.763742	1.000000	9.000000	

	Month	VisitorType	Weekend	Revenue	OperatingSystems	Browser	Region	TrafficType
count	12946	12946	12946	12946	12422.0	12946	12946	12946
unique	10	3	2	2	8.0	13	9	20
top	May	Returning_Visitor	False	False	2.0	2	1	2
freq	3533	11072	9929	10938	6673.0	8360	5031	4100

1. **Administrative_Duration**:

Nilai maksimum dari kolom durasi kunjungan administratif (Administrative_Duration) adalah 3398.75, sementara rata-rata (mean) hanya sekitar 80.37. Selisih antara nilai maksimum dan rata-rata yang sangat besar ini menunjukkan kemungkinan adanya outlier yang signifikan dalam data.

2. **Informational_Duration**:

Sama seperti kolom sebelumnya, nilai maksimum dari durasi kunjungan informatif (Informational_Duration) adalah 2549.375, sementara rata-ratanya hanya sekitar 34.14. Ini juga menunjukkan kemungkinan adanya outlier dalam data.

3. **ProductRelated**:

Nilai maksimum dari jumlah kunjungan terkait produk (ProductRelated) adalah 705, sementara rata-rata hanya sekitar 31.66. Sekali lagi, perbedaan yang signifikan antara nilai maksimum dan rata-rata menunjukkan kemungkinan adanya outlier dalam data.

4. **ProductRelated_Duration**:

Nilai maksimum dari durasi kunjungan terkait produk (ProductRelated_Duration) adalah 63973.522230, sementara rata-ratanya hanya sekitar 1192.74. Hal ini menunjukkan kemungkinan adanya outlier dalam data.

5. **PageValues**:

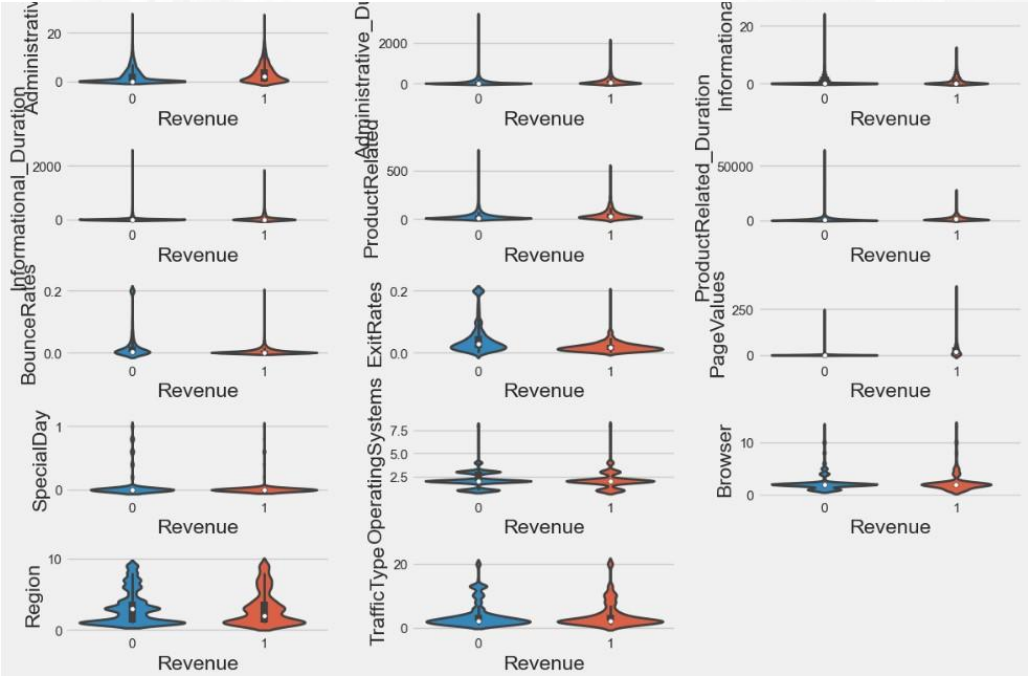
Nilai maksimum dari nilai halaman (PageValues) adalah 361.763742, yang jauh lebih tinggi dari rata-rata sekitar 5.88. Ini juga menunjukkan kemungkinan adanya outlier dalam data.

2. Univariate Analysis

Gunakan visualisasi untuk melihat distribusi masing-masing kolom (feature maupun target). Tuliskan hasil observasinya, misalnya jika ada suatu kolom yang distribusinya menarik (misal skewed, bimodal, ada outlier, ada nilai yang mendominasi, kategorinya terlalu banyak, dsb). Jelaskan juga apa yang harus di-follow up saat data pre-processing.

jawaban;
dugaan kami mengenai sebaran data yang skew diperkuat dengan hasil dari visualisasi dengan yang dapat dilihat di bawah ini:

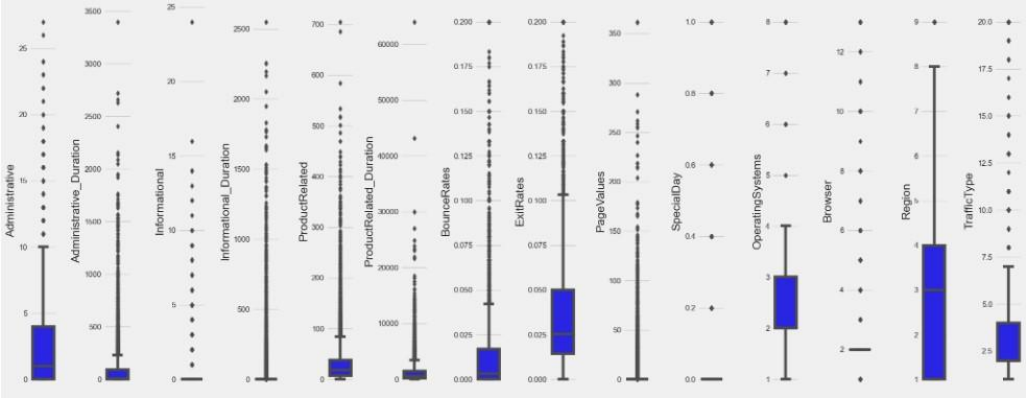
A. Violin Plot



Insight Violin Plot

Violin plot adalah salah satu jenis visualisasi yang memungkinkan kita untuk melihat distribusi dari sebuah fitur atau variabel. Ketika melihat violin plot untuk fitur 'Administrative_Duration', 'Informational', 'Informational_Duration', 'ProductRelated', 'ProductRelated_Duration', dan 'PageValues', kita melihat adanya beberapa titik di ujung-ujung plot yang berada jauh dari area utama plot atau gambaran utama distribusinya. Titik-titik ini menunjukkan adanya outlier.

B. Box Plot



Insight Box Plot

semua fitur box plot hanya menumpuk di bagian bawah, ini menunjukkan bahwa mayoritas data cenderung memiliki nilai yang rendah. Dalam konteks box plot, ini biasanya terlihat sebagai:

1. Median yang rendah:

Garis tengah pada kotak (box) akan cenderung berada pada bagian bawah plot, menunjukkan bahwa median nilai-nilai adalah rendah.

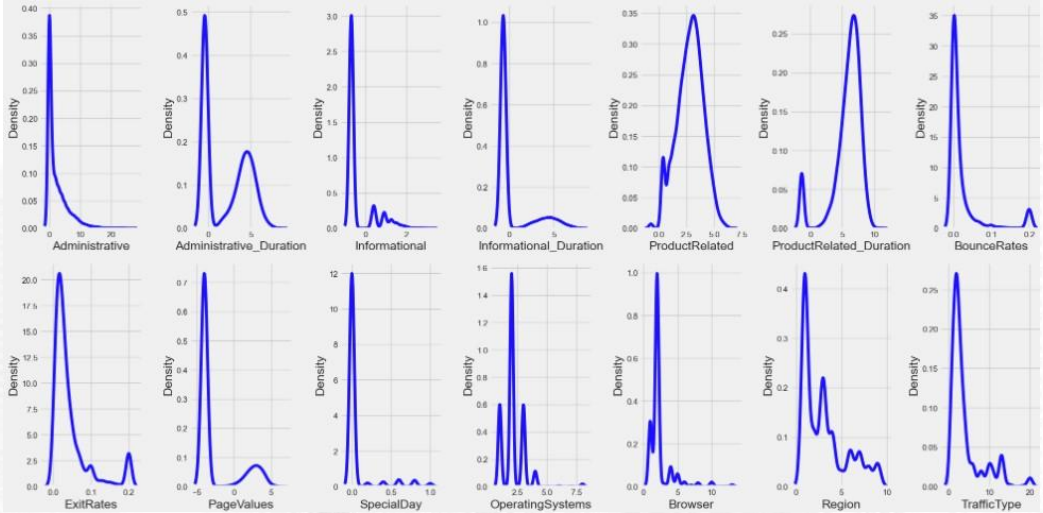
2. Kuartil Bawah (Q1) dan Kuartil Atas (Q3) yang rendah:

Bagian bawah dan bagian atas kotak juga akan cenderung berada pada bagian bawah plot, menunjukkan bahwa mayoritas nilai-nilai berada di bagian bawah rentang data.

3. 'Kumis' (whiskers) yang pendek:

'Kumis' atas dan 'kumis' bawah akan cenderung pendek atau mungkin tidak terlihat sama sekali, menunjukkan bahwa sebagian besar nilai-nilai berada dalam jarak yang relatif kecil dari median.

C. Sub Plot



Insight Sub Plot¶

1. Subplot di Kiri:

Ini adalah subplot pertama dalam satu baris. Fitur-fitur yang ditampilkan, seperti 'Administrative', 'Administrative_Duration', 'Informational', 'Informational_Duration', 'BounceRates', 'ExitRates', 'PageValues', dan 'SpecialDay', akan ditempatkan di sini. Dalam konteks "skew ke kiri", ini mengindikasikan bahwa distribusi data untuk fitur-fitur ini mungkin memiliki kecenderungan untuk condong ke kiri (skewed left), artinya nilai-nilai yang lebih kecil lebih umum daripada nilai-nilai yang lebih besar.

2. Subplot di Tengah:

Ini adalah subplot kedua dalam satu baris. Fitur 'ProductRelated' akan ditempatkan di sini. Dalam konteks "skew di tengah", tidak ada arti khusus yang terkait dengan istilah ini. Ini hanya menunjukkan bahwa fitur 'ProductRelated' ditempatkan di tengah antara subplot kiri dan kanan.

3. Subplot di Kanan:

Ini adalah subplot ketiga dalam satu baris. Fitur 'ProductRelated_Duration' akan ditempatkan di sini. Dalam konteks "skew ke kanan", ini menunjukkan bahwa distribusi data untuk fitur ini mungkin memiliki kecenderungan untuk condong ke kanan (skewed right), artinya nilai-nilai yang lebih besar lebih umum daripada nilai-nilai yang lebih kecil.

Follow up yang dapat dilakukan pada saat data preprocessing nanti adalah sebagai berikut:

1. Normalisasi Kolom dengan Distribusi Skewed:

Lakukan normalisasi pada kolom-kolom yang memiliki distribusi yang skew, seperti yang teridentifikasi sebelumnya, seperti 'Administrative_Duration', 'Informational', 'Informational_Duration', 'ProductRelated', 'ProductRelated_Duration', dan 'PageValues'. Hal ini dapat dilakukan dengan menggunakan teknik seperti log transformation atau Min-Max Scaling untuk mengurangi efek dari skewness.

2. Penanganan Data yang Hilang:

Identifikasi dan tangani kolom-kolom yang memiliki nilai kosong. Jika nilai kosong berada di bawah 10%, pilihan terbaik mungkin adalah menghapus baris atau kolom yang memiliki nilai kosong. Jika terdapat kolom yang memiliki lebih dari 10% nilai kosong, pertimbangkan untuk mengisi nilai kosong dengan menggunakan metode seperti mean, median, atau modus, tergantung pada distribusi data.

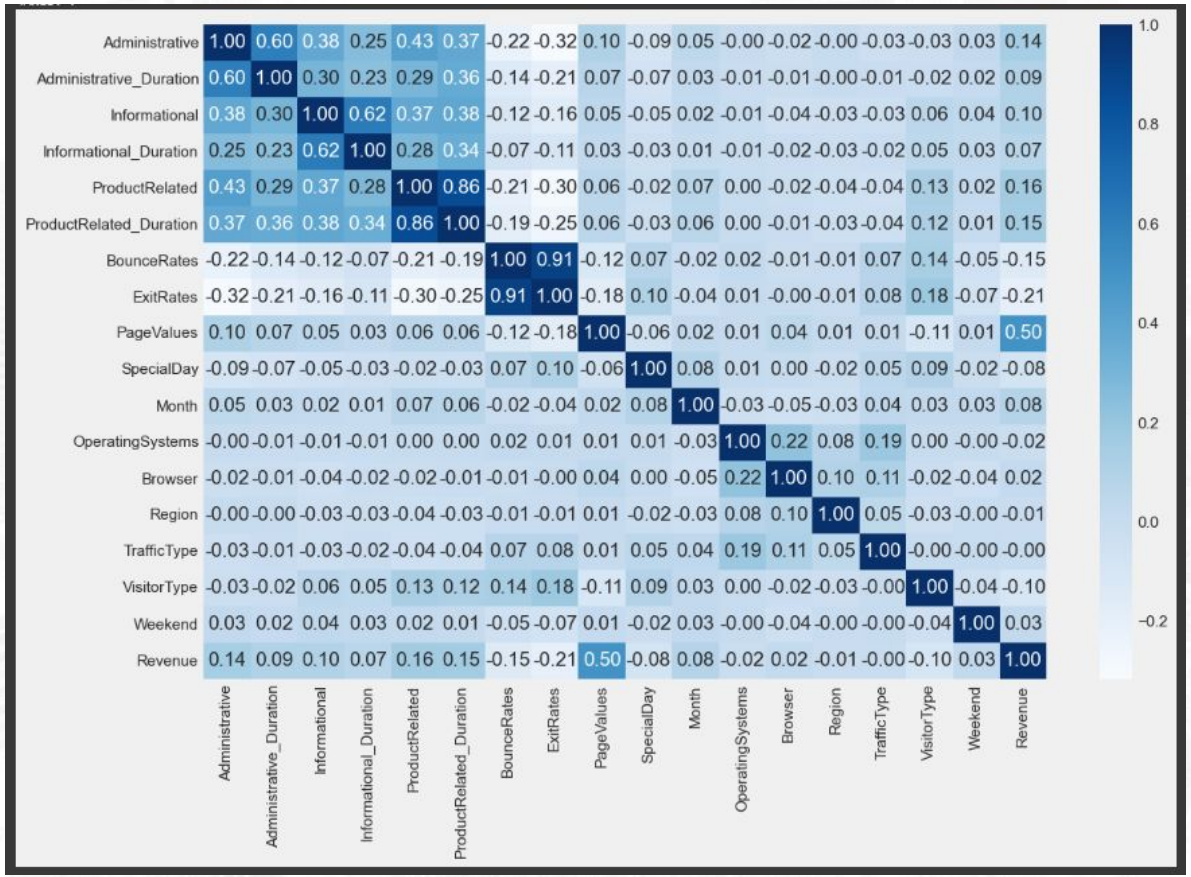
3. Penghapusan Data Duplikat:

Hapus data duplikat dari dataset untuk menghindari bias dan mendukung analisis yang akurat. Data duplikat dapat menyebabkan overfitting pada model dan mempengaruhi hasil analisis.

4. Normalisasi Outliers:

Identifikasi dan normalisasi outliers pada dataset. Outliers dapat mempengaruhi kualitas model yang dihasilkan dan menyebabkan distorsi dalam analisis. Metode normalisasi seperti winsorization atau penghapusan outliers dapat digunakan untuk menangani outliers.

3. Multivariate Analysis (15 poin)



Insight Heat Map

Heatmap adalah visualisasi yang berguna untuk menampilkan hubungan atau korelasi antara dua set data. Dalam konteks yang Anda berikan, heatmap sepertinya menampilkan korelasi antara fitur-fitur tertentu dengan fitur 'revenue'. Mari kita jelaskan setiap poin yang Anda sebutkan:

1. PageValues dan Revenue memiliki nilai paling tinggi pertama yaitu 0.50:

Ini menunjukkan bahwa ada korelasi positif yang kuat antara fitur PageValues (nilai halaman) dan Revenue. Artinya, semakin tinggi nilai PageValues, semakin besar kemungkinan pengunjung akan menghasilkan pendapatan.

2. ProductRelated dan Revenue memiliki nilai tinggi ke 2 yaitu 0.16:

Ini menunjukkan bahwa ada korelasi positif antara fitur ProductRelated (kunjungan terkait produk) dan Revenue, meskipun korelasinya tidak sekuat dengan PageValues.

3. ProductRelated Duration dan Revenue memiliki nilai tinggi ke 3 yaitu 0.16:

Korelasi positif yang sama dengan ProductRelated juga terjadi antara ProductRelated Duration (durasi kunjungan terkait produk) dan Revenue.

4. Exit Rate dan Revenue memiliki nilai paling rendah yaitu -0.21:

Ini menunjukkan bahwa ada korelasi negatif yang cukup kuat antara Exit Rate (tingkat keluar) dan Revenue. Artinya, semakin tinggi tingkat keluar, semakin kecil kemungkinan pengunjung akan menghasilkan pendapatan.

5. Bounce Rate dan Revenue memiliki nilai rendah ke 2 yaitu -0.15:

Korelasi negatif juga terjadi antara Bounce Rate (tingkat pentalan) dan Revenue. Semakin tinggi tingkat pentalan, semakin kecil kemungkinan pengunjung akan menghasilkan pendapatan.

6. Visitor Type dan Revenue memiliki nilai rendah ke 3 yaitu -0.10:

Korelasi negatif antara Visitor Type (tipe pengunjung) dan Revenue menunjukkan bahwa jenis pengunjung tertentu mungkin memiliki dampak negatif terhadap pendapatan.

A. Bagaimana korelasi antara masing-masing feature dan label. Kira-kira feature mana saja yang paling relevan dan harus dipertahankan?

Setelah dilakukan visualisasi dengan heatmap kami dapat menyimpulkan beberapa korelasi kolom fitur terhadap kolom target kami (revenue) sebagai berikut:

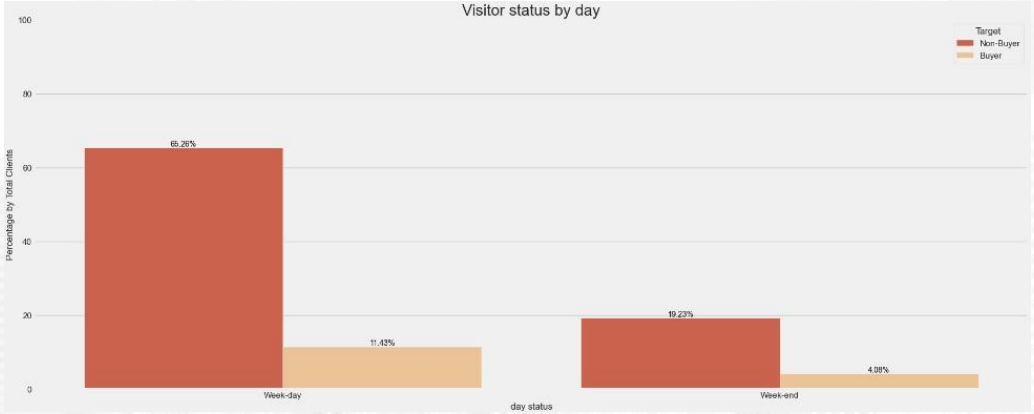
- korelasi positif terjadi dengan kolom PageValue (paling kuat), Product_Related, ProductRelated_Duration, Administrative, Informational, Administrative_Duration, Informational_duration, Month.
- sedangkan korelasi negatif terjadi dengan kolom VisitorType, SpecialDay, ExitRates, BounceRates .
- Serta korelasi yang lemah atau bahkan tidak ada korelasi sama sekali terjadi pada kolom Region, TrafficType, Browser, OperatingSystems, Weekend.

B. Bagaimana korelasi antar-feature, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap feature itu?

- selain itu korelasi positif kuat juga terjadi antara ProductRelated_Duration - ProductRelated, Informational_Duration - Informational, Product_Related - Administrative, serta Administrative - Administrative_Duration.
- korelasi negatif juga terjadi pada kolom ExitRates - Product_Related, BounceRates - ProductRelated.
- Selain itu korelasi antara kolom ProductRelated_duration - Product_related, dan ExitRates - BounceRates memiliki nilai di atas 0.7 yang memungkinkan antara kolom tersebut terjadi redundansi
- beberapa kolom tidak memiliki korelasi atau korelasinya sangat kecil, diantaranya region - administrative, OperatingSystems - SpecialDay, dan lain-lain

4. Business Insight

A. Menambah Fitur Weekend status Berdasarkan Weekend



Insight Weekend Status VS Target

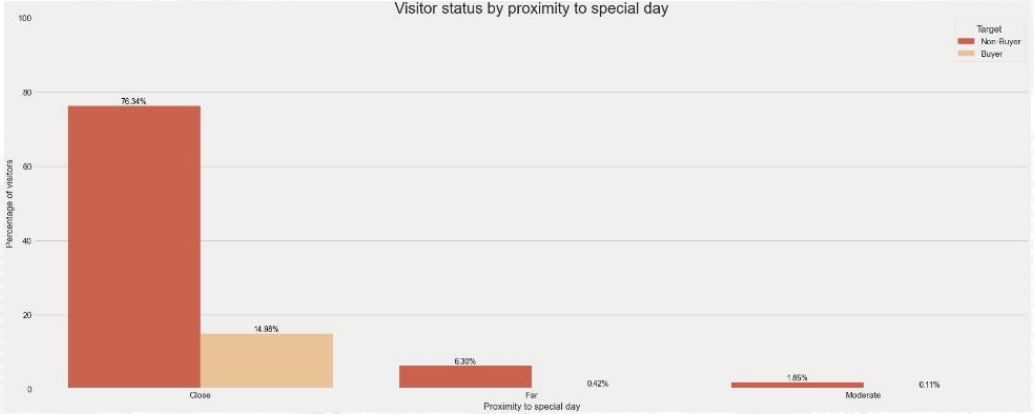
Business Insight:

1. Polap Pembelian Berdasarkan Hari dalam Seminggu:
2. Potensi Pembelian pada Akhir Pekan:

Recommendation:

1. Optimalkan Strategi Pemasaran pada Akhir Pekan:
2. Personalisasi Pengalaman Pembelian pada Hari Kerja:

B. Menambah Fitur Special Day Category Berdasarkan Special Day



Insight Special Day Category VS Target

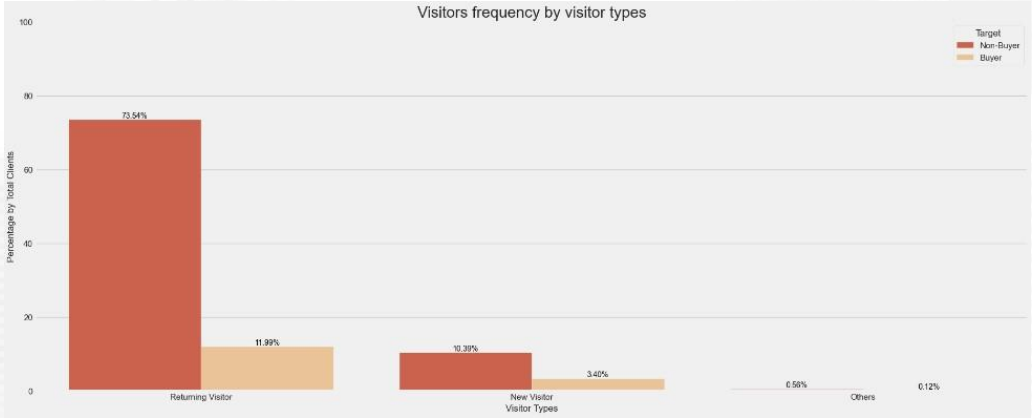
Business Insight:

1. Pengaruh Hari Khusus pada Perilaku Pembelian:

Recommendation:

1. Penyesuaian Strategi Pemasaran
2. Peningkatan Keterlibatan pada Hari Khusus

C. Insight Menambah Fitur Visitor Types Berdasarkan Target



Insight Fitur Visitor Types VS Target

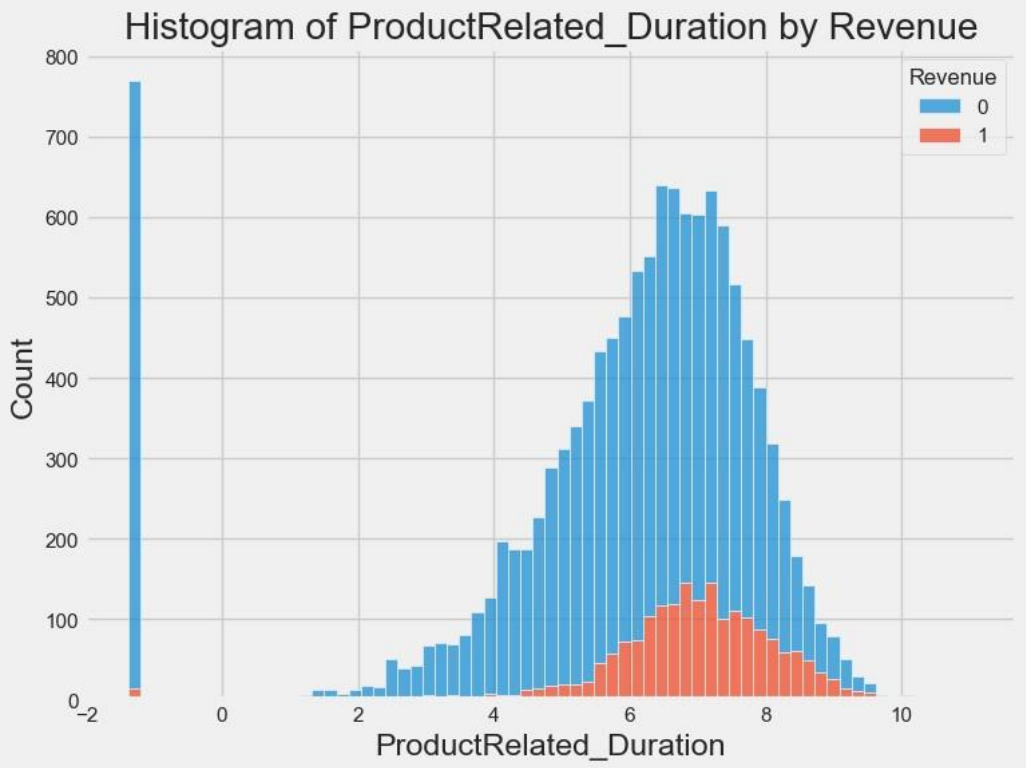
Business Insight:

1. Polap Pembelian Berdasarkan Tipe Pengunjung:
2. Perilaku Pembelian dan Retensi Pelanggan:

Recommendation:

1. Personalisasi Pengalaman Pengunjung
2. Strategi Retensi Pelanggan
3. Peningkatan Pengalaman Pengguna Baru
4. Analisis Lebih Lanjut terhadap Pengunjung Kategori 'Others'

D. Product Related Duration VS Revenue



Insight Product Related Duration VS Revenue

Jika durasi yang dihabiskan oleh pengguna pada halaman terkait produk ('Product Related Duration') memiliki skew ke kanan (artinya distribusinya condong ke nilai-nilai yang lebih tinggi) dan jumlah pendapatan ('Revenue') masih jauh dari merata antara kategori 'Buyer' dan 'Non-Buyer', ini mengindikasikan adanya potensi untuk meningkatkan konversi pembelian.

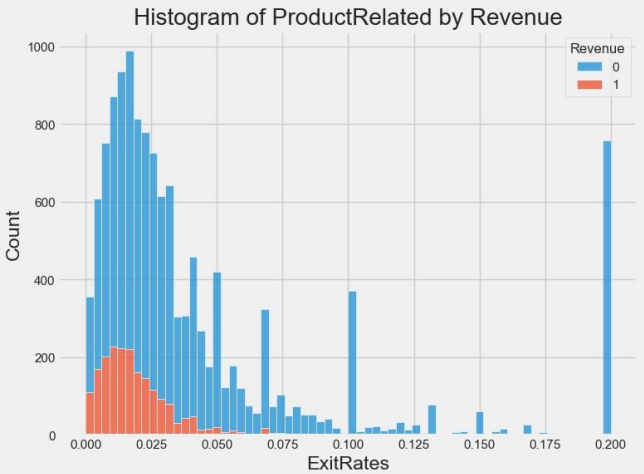
Business Insight:

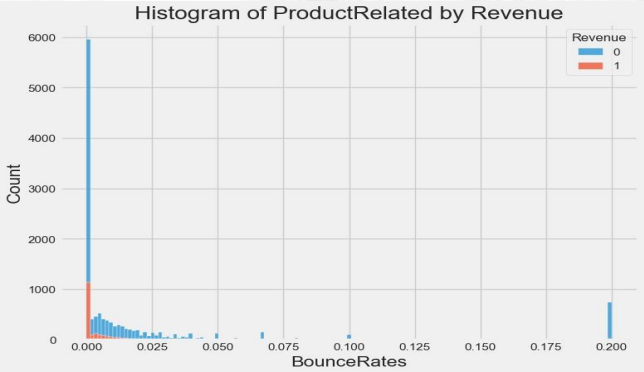
1. Durasi yang dihabiskan oleh pengguna pada halaman terkait produk ('Product Related Duration') yang condong ke nilai yang lebih tinggi menunjukkan minat yang kuat dari pengguna terhadap produk atau layanan yang ditawarkan.
2. Namun, meskipun minat ini tinggi, jumlah pendapatan ('Revenue') dari pengguna yang melakukan pembelian ('Buyer') masih jauh lebih rendah daripada yang tidak ('Non-Buyer').

Recommendation:

1. Optimalkan Pengalaman Pengguna
2. Personalisasi dan Rekomendasi Produk

E. Exit Rates & Bounce Rate VS Revenue





Insight Exit Rates & Bounce Rate VS Revenue

Jika tingkat keluar dari halaman produk ('Exit Rate & Bounce Rate') memiliki skew ke kiri (artinya distribusinya condong ke nilai-nilai yang lebih rendah) dan jumlah pendapatan ('Revenue') masih jauh lebih rendah dari yang diharapkan, ini mengindikasikan adanya potensi masalah dalam proses pembelian atau pengalaman pengguna yang perlu diperbaiki.

Business Insight:

1. Tingkat keluar yang rendah dari halaman produk ('Exit Rate & Bounce Rate') menunjukkan bahwa pengguna cenderung tinggal lebih lama di halaman tersebut, yang seharusnya meningkatkan kemungkinan pembelian.
2. Namun, meskipun pengguna tinggal lebih lama, jumlah pendapatan ('Revenue') masih jauh lebih rendah dari yang diharapkan, menunjukkan bahwa ada masalah dalam mengubah minat pengguna menjadi tindakan pembelian.

Recommendation:

1. Analisis dan Perbaikan Proses Pembelian
2. Personalisasi dan Rekomendasi Produk
3. Analisis Pelanggan yang Telah Keluar

5. Git

Link Github : https://github.com/agussetiana/final_project_rakamin_goingfast