Rakamin Academy

Homework EDA

Final Project - Stage 1

ONLINE SHOPPERS PURCHASING INTENTION

	Administrative	Administrative_Duration	Informational	Informational_Duration	ProductRelated	ProductRelated_Duration	BounceRates	ExitRates	PageValues	SpecialDay	Month	OperatingSystems	Browser	Region	TrafficType	VisitorType	Weekend	Revenue
0	0	0.0	0	0.0	1	0.000000	0.200000	0.200000	0.0	0.0	Feb	1	1	1	1	Returning_Visitor	False	False
1	0	0.0	0	0.0	2	64.000000	0.000000	0.100000	0.0	0.0	Feb	2	2	1	2	Returning_Visitor	False	False
2	0	0.0	0	0.0	1	0.000000	0.200000	0.200000	0.0	0.0	Feb	4	1	9	3	Returning_Visitor	False	False
3	0	0.0	0	0.0	2	2.666667	0.050000	0.140000	0.0	0.0	Feb	3	2	2	4	Returning_Visitor	False	False
4	0	0.0	0	0.0	10	627.500000	0.020000	0.050000	0.0	0.0	Feb	3	3	1	4	Returning_Visitor	True	False
5	0	0.0	0	0.0	19	154.216667	0.015789	0.024561	0.0	0.0	Feb	2	2	1	3	Returning_Visitor	False	False
6	0	0.0	0	0.0	1	0.000000	0.200000	0.200000	0.0	0.4	Feb	2	4	3	3	Returning_Visitor	False	False
7	1	0.0	0	0.0	0	0.000000	0.200000	0.200000	0.0	0.0	Feb	1	2	1	5	Returning_Visitor	True	False
8	0	0.0	0	0.0	2	37.000000	0.000000	0.100000	0.0	0.8	Feb	2	2	2	3	Returning_Visitor	False	False
9	0	0.0	0	0.0	3	738.000000	0.000000	0.022222	0.0	0.4	Feb	2	4	1	2	Returning_Visitor	False	False

1. Descriptive Statistics

Rakamin Academy

Gunakan function **`info`** dan **`describe`** pada dataset final project kalian. Tuliskan hasil observasinya, seperti:

A. Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?
B. Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?
C. Apakah ada kolom yang memiliki nilai summary agak aneh? (min/mean/median/max/unique/top/freq)

*Untuk masing-masing jenis observasi, tuliskan juga jika tidak ada masalah, misal untuk A: “Semua tipe data sudah sesuai”

Descriptive Statistics

Rakamin Academy

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12330 entries, 0 to 12329
Data columns (total 18 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   Administrative           12330 non-null  int64
 1   Administrative_Duration  12330 non-null  float64
 2   Informational            12330 non-null  int64
 3   Informational_Duration   12330 non-null  float64
 4   ProductRelated           12330 non-null  int64
 5   ProductRelated_Duration  12330 non-null  float64
 6   BounceRates              12330 non-null  float64
 7   ExitRates                12330 non-null  float64
 8   PageValues               12330 non-null  float64
 9   SpecialDay               12330 non-null  float64
 10  Month                    12330 non-null  object
 11  OperatingSystems         12330 non-null  int64
 12  Browser                  12330 non-null  int64
 13  Region                   12330 non-null  int64
 14  TrafficType              12330 non-null  int64
 15  VisitorType              12330 non-null  object
 16  Weekend                  12330 non-null  bool
 17  Revenue                  12330 non-null  bool
dtypes: bool(2), float64(7), int64(7), object(2)
memory usage: 1.5+ MB
```

```
df[nums].describe().T
```

	count	mean	std	min	25%	50%	75%	max
Administrative	12330.0	2.315166	3.321784	0.0	0.000000	1.000000	4.000000	27.000000
Administrative_Duration	12330.0	80.818611	176.779107	0.0	0.000000	7.500000	93.256250	3398.750000
Informational	12330.0	0.503569	1.270156	0.0	0.000000	0.000000	0.000000	24.000000
Informational_Duration	12330.0	34.472398	140.749294	0.0	0.000000	0.000000	0.000000	2549.375000
ProductRelated	12330.0	31.731468	44.475503	0.0	7.000000	18.000000	38.000000	705.000000
ProductRelated_Duration	12330.0	1194.746220	1913.669288	0.0	184.137500	598.936905	1464.157214	63973.522230
BounceRates	12330.0	0.022191	0.048488	0.0	0.000000	0.003112	0.016813	0.200000
ExitRates	12330.0	0.043073	0.048597	0.0	0.014286	0.025156	0.050000	0.200000
PageValues	12330.0	5.889258	18.568437	0.0	0.000000	0.000000	0.000000	361.763742

```
df[cats1].describe().T
```

	count	mean	std	min	25%	50%	75%	max
SpecialDay	12330.0	0.061427	0.198917	0.0	0.0	0.0	0.0	1.0
OperatingSystems	12330.0	2.124006	0.911325	1.0	2.0	2.0	3.0	8.0
Browser	12330.0	2.357097	1.717277	1.0	2.0	2.0	2.0	13.0
Region	12330.0	3.147364	2.401591	1.0	1.0	3.0	4.0	9.0
TrafficType	12330.0	4.069586	4.025169	1.0	2.0	2.0	4.0	20.0

```
df[cats2].describe()
```

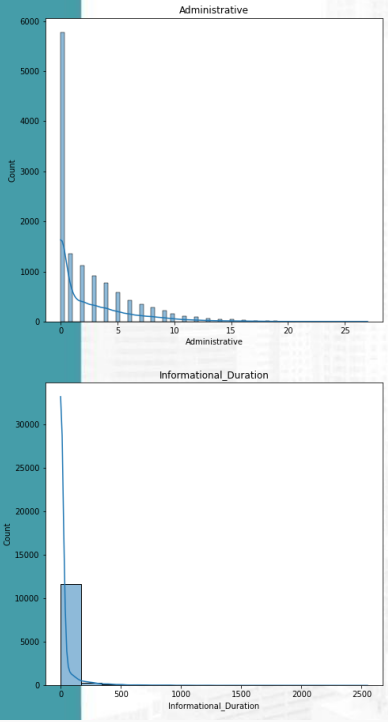
	Month	VisitorType	Weekend	Revenue
count	12330	12330	12330	12330
unique	10	3	2	2
top	May	Returning_Visitor	False	False
freq	3364	10551	9462	10422

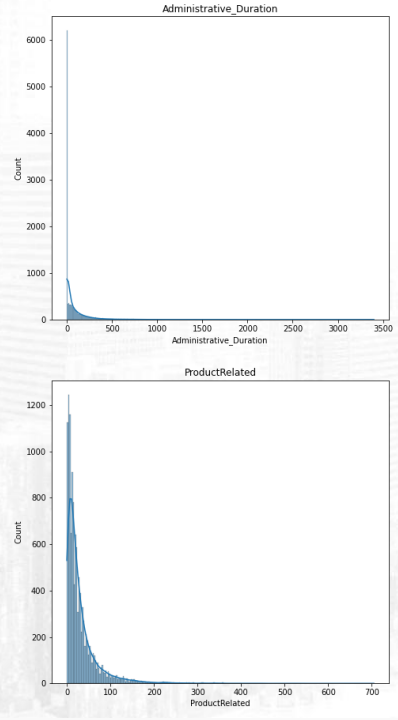
- Tipe data cats1 [SpecialDay, OperatingSystems, Browser, Region, TrafficType] adalah int dan float. Tipe data cats1 tidak sesuai untuk categorical feature, lebih baik tipe data diganti obj untuk mempermudah pembuatan model Machine Learning
- Pada dataset ini tidak ada nilai null dan arti nilai '0' bukan berarti nilai null
- Setiap feature pada dataset tidak memiliki nilai summary yang aneh

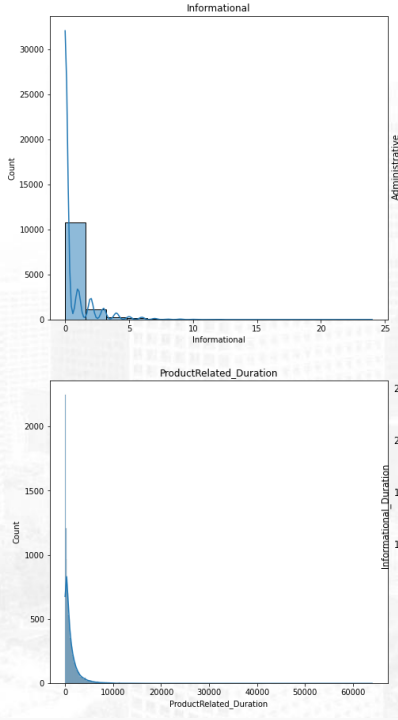
2. Univariate Analysis

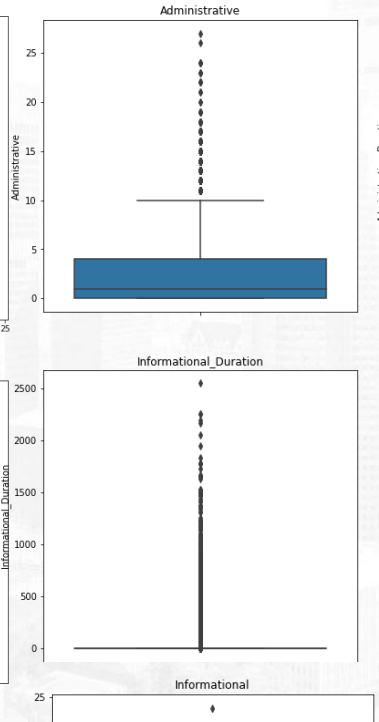
Gunakan visualisasi untuk melihat distribusi masing-masing kolom (feature maupun target). Tuliskan hasil observasinya, misalnya jika ada suatu kolom yang distribusinya menarik (misal skewed, bimodal, ada outlier, ada nilai yang mendominasi, kategorinya terlalu banyak, dsb). Jelaskan juga apa yang harus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

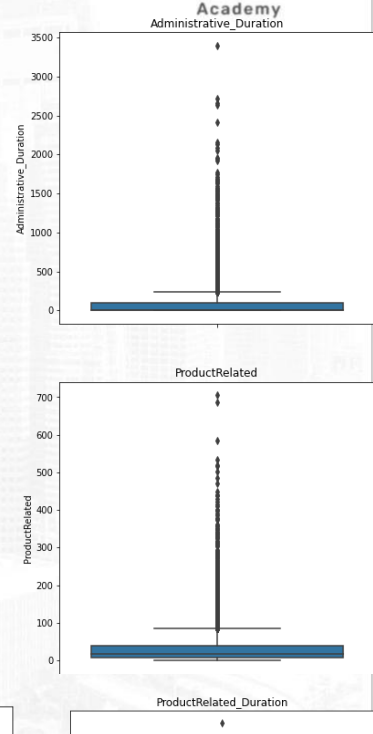
Univariate Analysis – Numerical Features

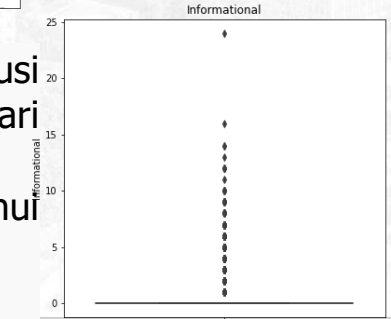


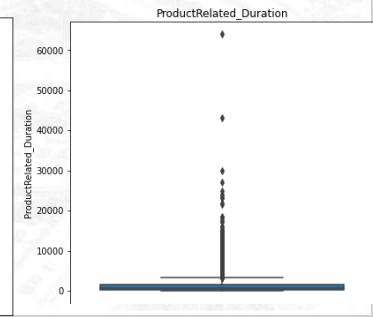






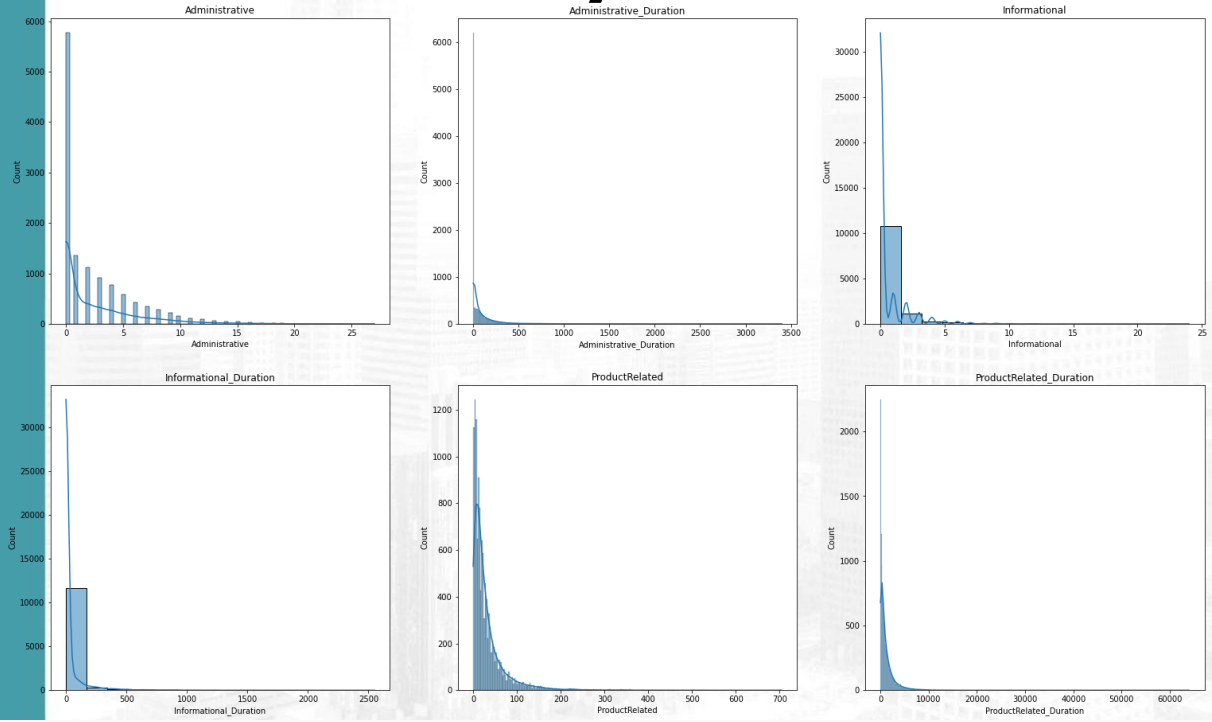


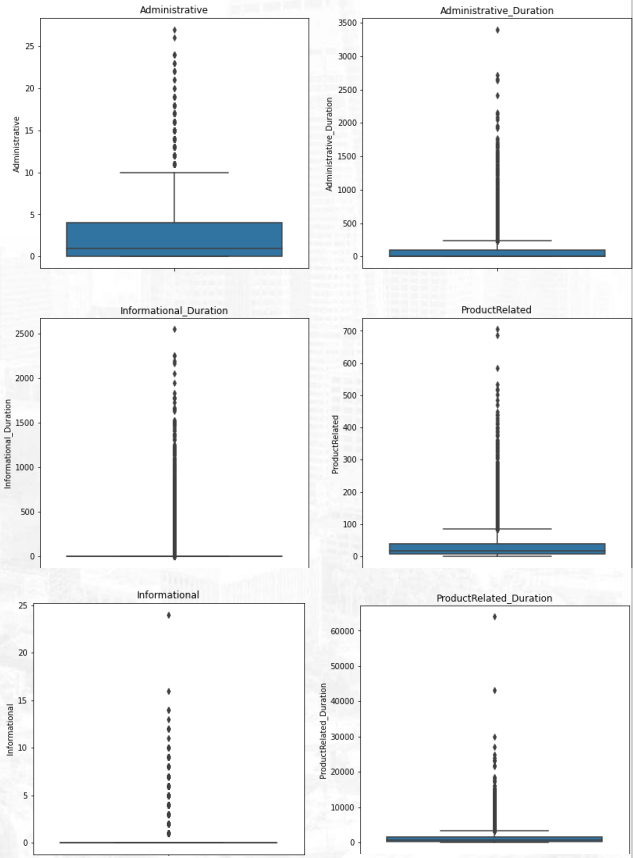




- Berdasarkan histplos dan boxplot, numerical features memiliki distribusi yang skew positif dan tidak bimodal, selain itu jika dilihat dari boxplotnya semua numerical features memiliki outlier besar.
- Perlu dilakukan identifikasi outlier lebih lanjut untuk mengetahui handling outlier tersebut dengan metode yang cocok untuk model ML
- Data yang skewed dapat dihandling dengan log transformation

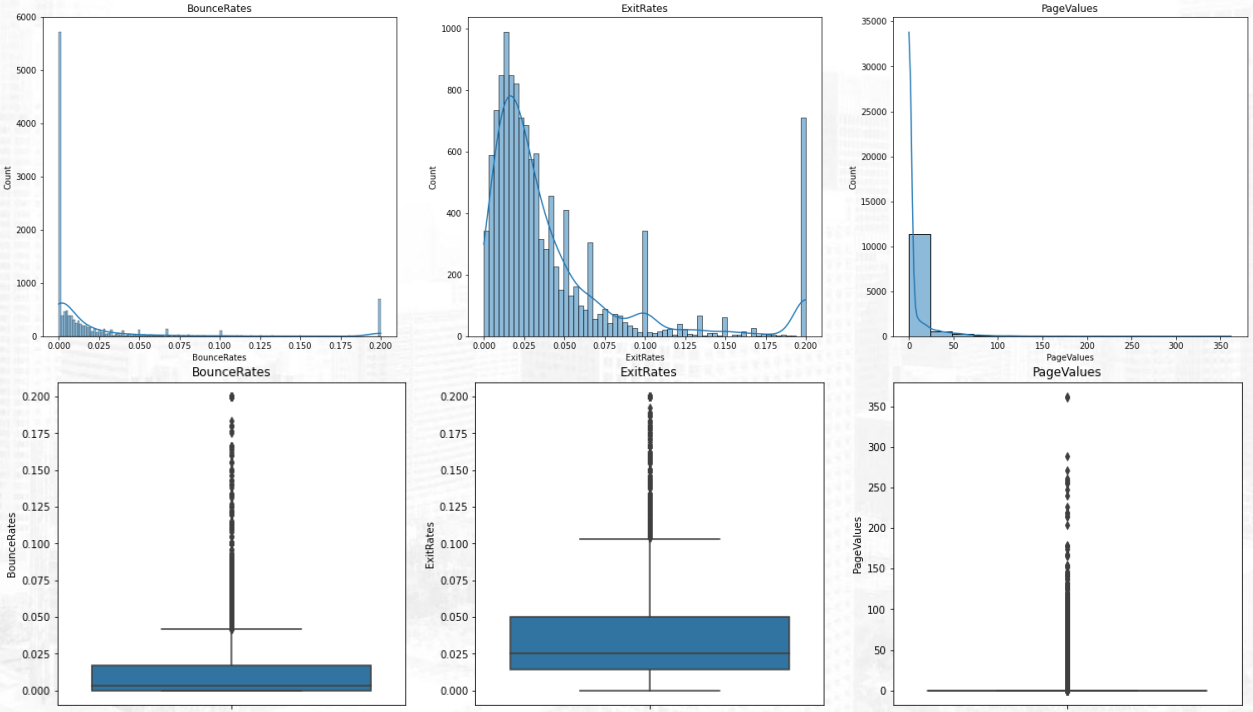
Univariate Analysis – Numerical Features





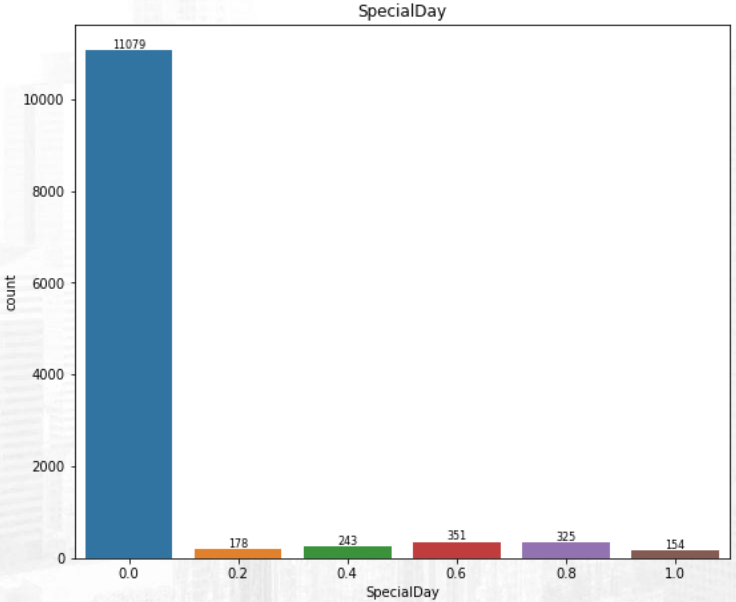
- Banyak user yang mengunjungi Administrative dan Informational pages dilihat dari jumlah user yang berdistribusi tinggi pada number of pages 0
- Karena banyaknya user yang mengunjungi Administrative dan Informational pages hanya beberapa halaman awal saja, user cenderung tidak banyak menghabiskan waktu pada kedua pages tersebut dan distribusi durasi kunjungan terbanyak pada detik 0.
- Banyak user yang mengunjungi Product Related pages setidaknya 1 halaman, namun ada juga beberapa user yang tidak mengunjungi atau bahkan mengunjungi Product Related pages sebanyak diatas 100 halaman. Durasi kunjungan Product Related pages antara 0 sampai dengan 100 detik.

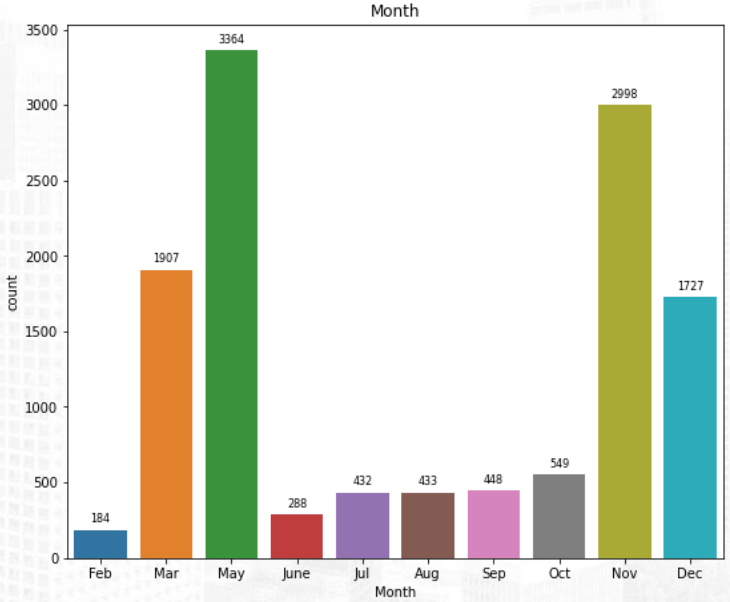
Univariate Analysis – Numerical Features



- Bounce rates dari histplot diatas dapat menunjukkan bahwa kebanyakan orang bahkan tidak mengunjungi, dan hanya sedikit orang mengunjungi dan pindah halaman
- Banyak pengunjung keluar dari halaman Exit rates berkisaran pada nilai 25-50 detik, namun ada juga yang mencapai 200 detik
- Page value rendah dikarenakan banyaknya user yang hanya mengunjungi sedikit pages (hanya di halaman awal) dan tidak melanjutkan ke halaman transaksi dari shopping site kita

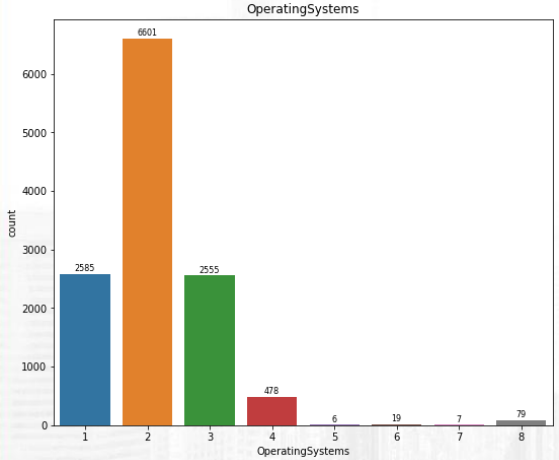
Univariate Analysis – Categorical Features

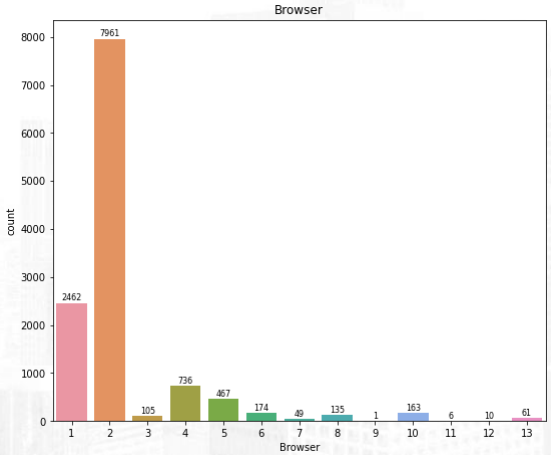


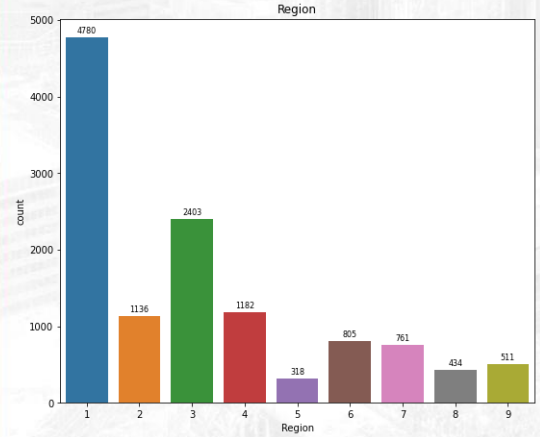


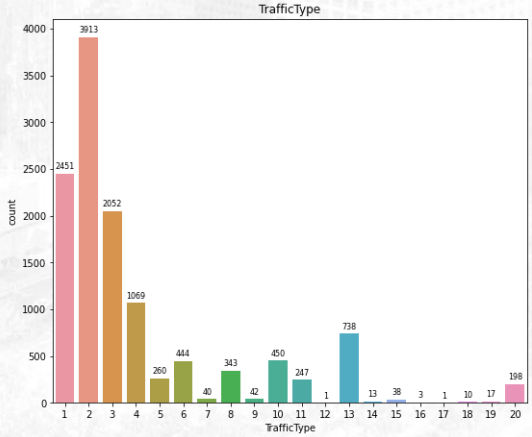
- Sebanyak 89.9% user lebih dominan berkunjung di waktu yang tidak dekat dengan Special Day
- Jumlah user terbanyak yang mengunjungi shopping site terdapat pada bulan Mei (27.28%), disusul dengan bulan November (24.32%) dan Maret (15.47%). Namun pada bulan Februari memiliki pengunjung yang rendah walaupun dalam bulan Februari terdapat special days (Valentine's day). Bulan Januari dan April tidak termasuk ke dalam dataset ini. Untuk selanjutnya dalam perekaman dan perekapan data perlu dilakukan secara lengkap setiap waktu atau bulannya agar hasil analisis dan model yang kita buat tepat dan terhindar dari kesalahan pengambilan keputusan.

Univariate Analysis – Categorical Features



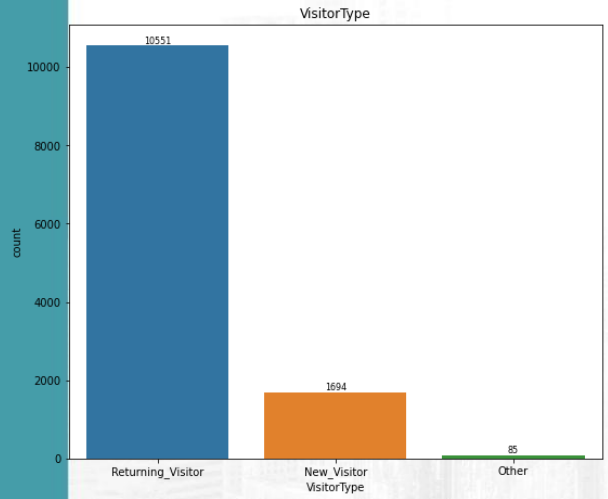


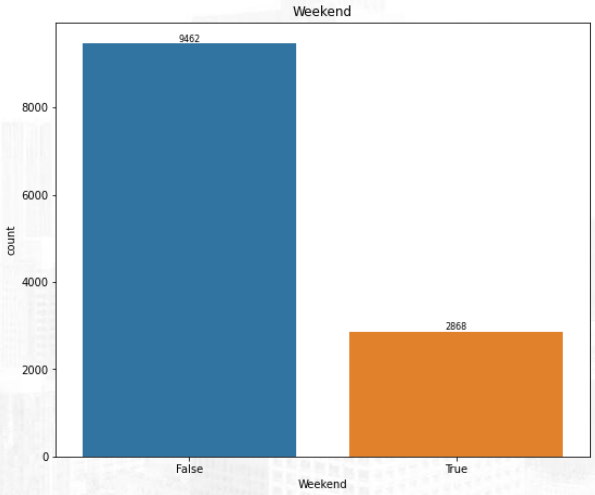


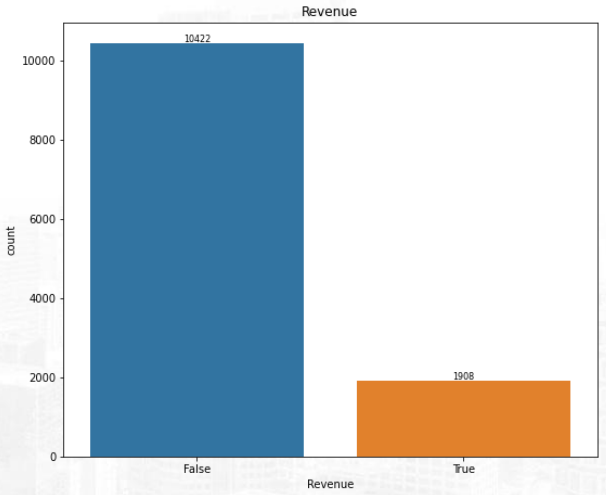


- OS 2 merupakan operating system yang lebih banyak digunakan oleh user dalam mengunjungi shopping site, disusul dengan OS 1 dan 3.
- Browser 2 merupakan browser yang paling banyak digunakan oleh user dalam mengunjungi shopping site sebesar 64.57% user yang berkunjung, disusul dengan Browser 1 dan 4.
- Region 1 memiliki jumlah user yang paling banyak mengunjungi shopping site yaitu sebesar 38.77% user yang berkunjung, disusul dengan Region 3 dan 4.
- Traffic Type 2 merupakan sumber traffic yang paling banyak digunakan untuk user mengunjungi shopping site, sebanyak 31.74% user yang berkunjung, kemudian disusul dengan sumber traffic type 1 dan 3.

Univariate Analysis – Categorical Features







- Sebanyak 85.57% Returning visitor merupakan user yang paling banyak mengunjungi shopping site.
- Sebanyak 76.74% user melakukan transaksi pada di hari selain Weekend.
- Hanya 15.47% atau 1908 user yang berkunjung dan melakukakan transaksi. Dikarena feature target (Revenue) memiliki persentase data 84.53% false dan hanya 15.47% yang true, sehingga dasatet ini dapat dikatakan sebagai imbalance data. Oleh karena itu perlu diatasi dengan metode Class Weight.

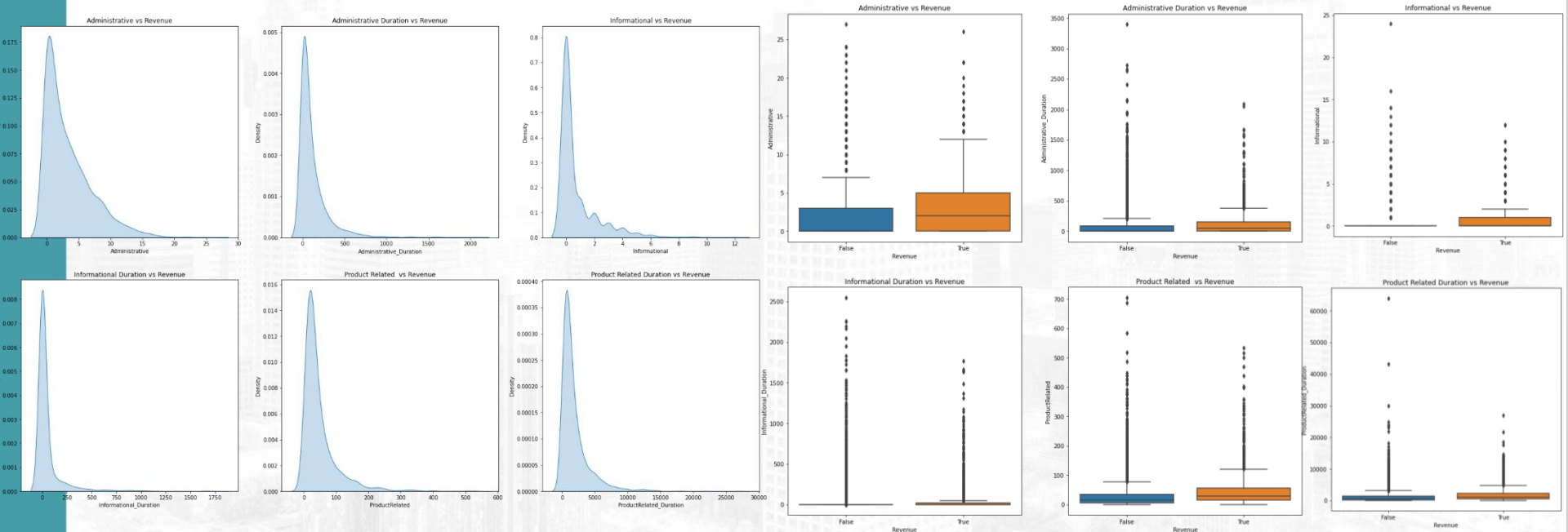
3. Multivariate Analysis

Lakukan multivariate analysis (seperti correlation heatmap dan category plots, sesuai yang diajarkan di kelas). Tuliskan hasil observasinya, seperti:

A. Bagaimana korelasi antara masing-masing feature dan label. Kira-kira feature mana saja yang paling relevan dan harus dipertahankan?
B. Bagaimana korelasi antar-feature, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap feature itu?

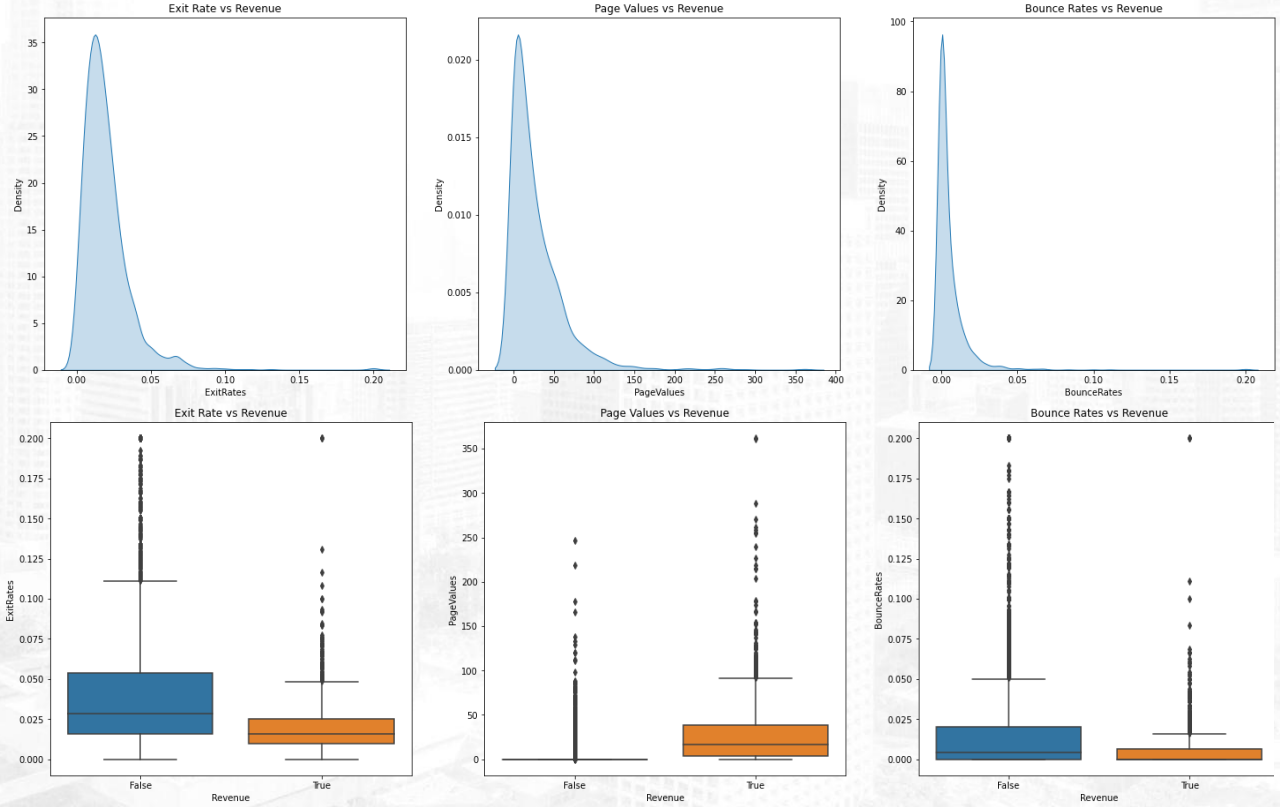
*Tuliskan juga jika memang tidak ada feature yang saling berkorelasi

Bivariate Analysis – Numerical Features



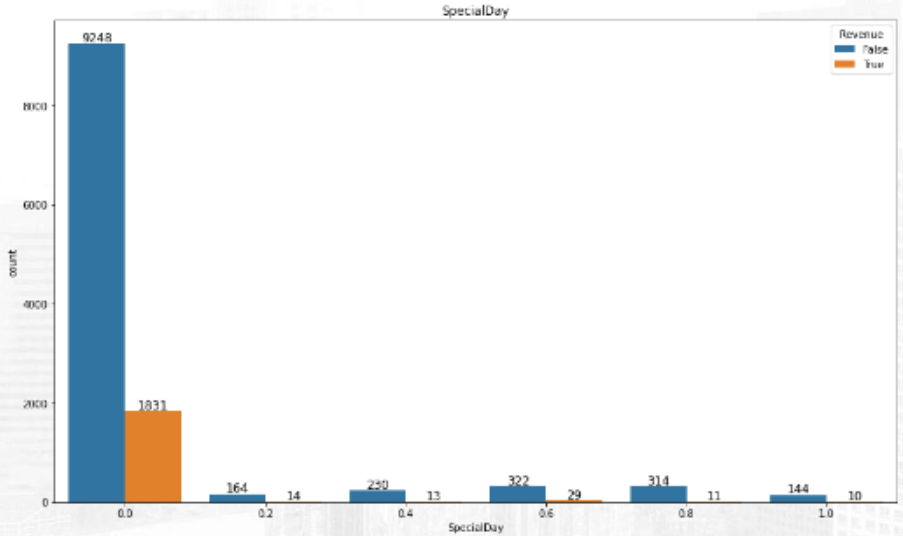
- Pada Administrative dan Administrative_Duration menunujukkan bahwa setiap kali pengunjung yang mengunjungi halaman bukan berarti mereka membeli.
- Pada Informasi dan Information_Duration menunjukkan bahwa orang menghabiskan waktu membaca halaman informasi untuk memutuskan membeli sesuatu atau tidak. Dan kebanykan pengunjung hanya berkunjung dan tidak membeli.
- ProductRelated dan ProductRelated_Duration menunjukkan banyak user mengunjungi halaman terkait produk dan menghabiskan waktu disana dengan jumlah paling banyak mengunjungi halaman terkait sebesar 0-400.

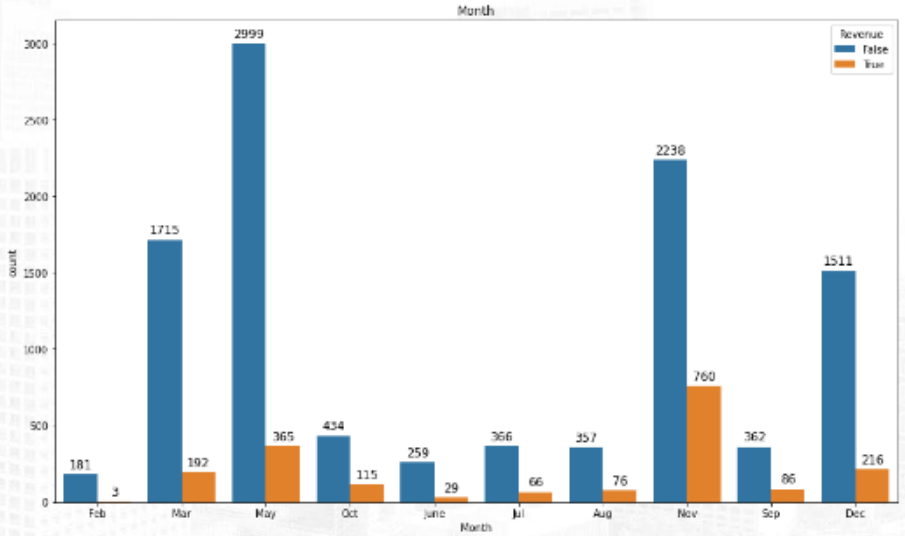
Bivariate Analysis – Numerical Features



- Tingginya angka BouncesRates dan ExitRates menunjukkan bahwa kedua fitur tersebut tidak menghasilkan Revenue
- Angka PageValues yang menghasilkan Revenue lebih dominan dikarenakan PageValues adalah rata-rata kunjungan halaman keranjang (cart) atau halaman checkout

Bivariate Analysis – Categorical Features

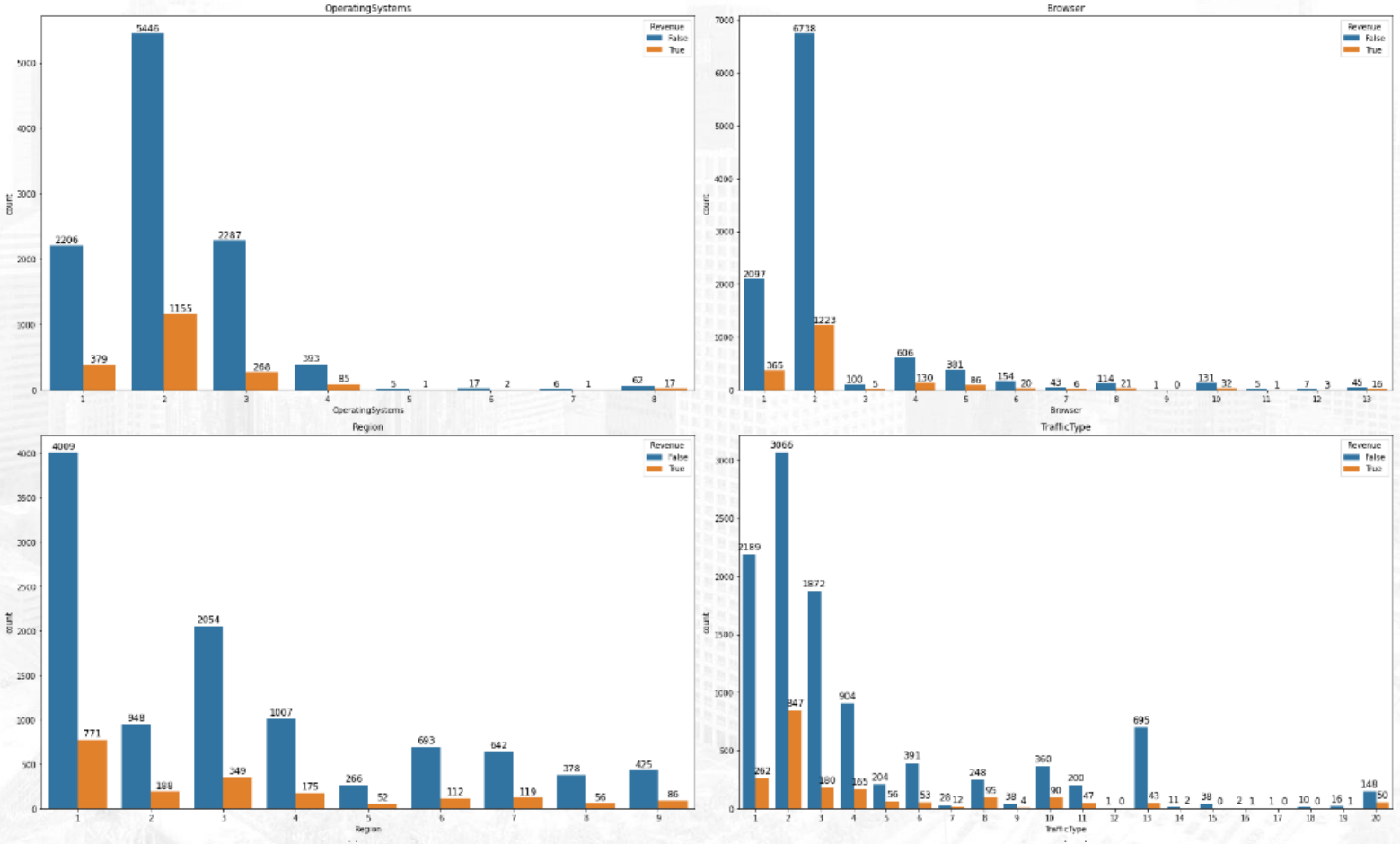




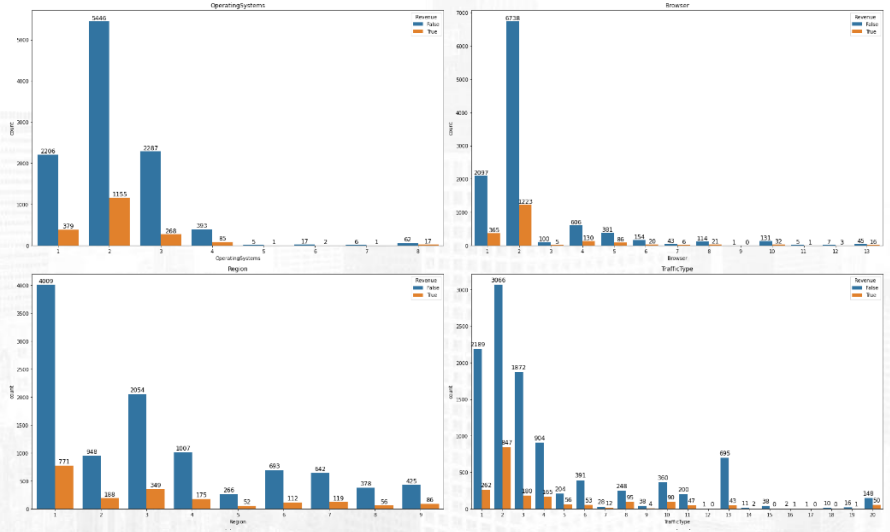
- Persentase pengunjung pada feature Special day yang banyak melakukan transaksi (Revenue) pada 0.0 sebanyak 16.53% sedangkan pada 1.0 ada 6.49% banyak pengunjung yang melakukan transaksi. Dapat dilihat dari persentase diatas daya minat beli user lebih condong sebelum Special day.
- Jumlah user yang paling banyak melakukan transaksi pada bulan November sebesar 25.35%, bulan Oktober 20.95%, dan bulan September 19.20%. Namun pada bulan Mei paling banyak pengunjung yang tidak melakukan transaksi.

Bivariate Analysis – Categorical Features

Rakamin Academy

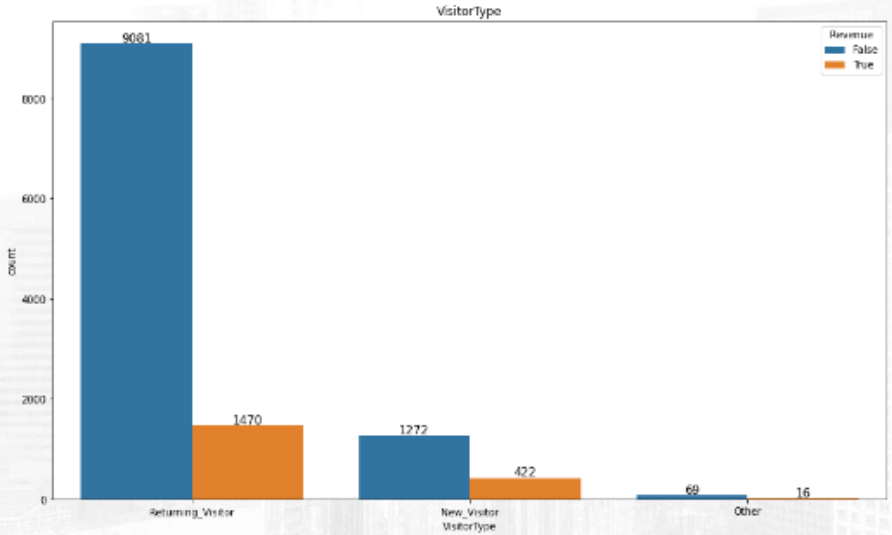


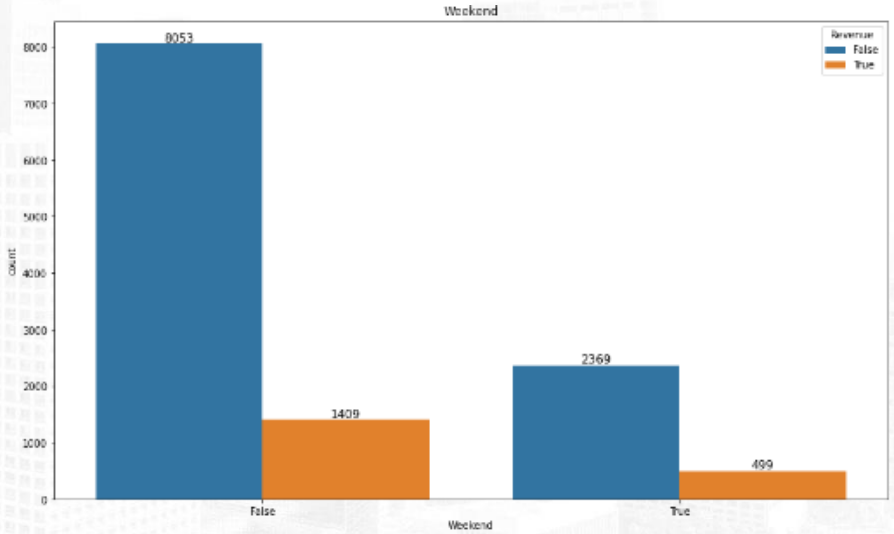
Bivariate Analysis – Categorical Features



- Pada feature OperatingSystem jumlah pengunjung yang banyak melakukan transaksi pada OS 4 sebesar 17.78% dan pada OS 2 sebesar 17.50%.
- Jumlah user yang paling banyak melakukan transaksi terdapat pada browser 12 sebesar 30.00% dan browser 13 sebesar 26.23%.
- Jumlah user terbanyak melakukan transaksi terdapat pada region 9 sebesar 16.83%, region 2 sebesar 16.55%, dan region 5 sebesar 16.35%.
- TrafficType 7 terdapat 30.00% pengunjung yang banyak melakukan transaksi dan traffictype 8 sebesar 27.70%.

Bivariate Analysis – Categorical Features

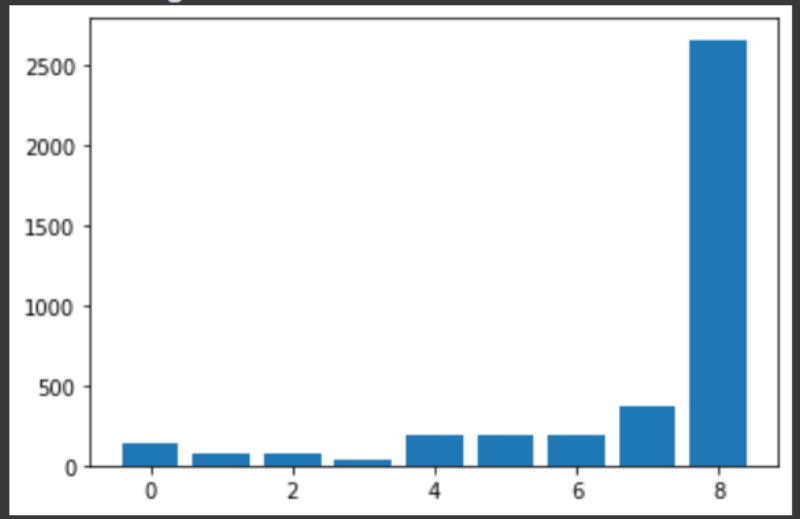




- Banyak new visitor yang melakukan transaksi sebesar 24.91%
- Jumlah pengunjung yang melakukan transaksi pada weekend sebesar 17.40%

Multivariate Analysis – Numerical Correlation

```
Feature Administrative: 143.684475
Feature Administrative_Duration: 75.006355
Feature Informational: 68.224184
Feature Informational_Duration: 38.486448
Feature ProductRelated: 195.789853
Feature ProductRelated_Duration: 184.589367
Feature BounceRates: 194.402926
Feature ExitRates: 373.064180
Feature PageValues: 2656.238864
```

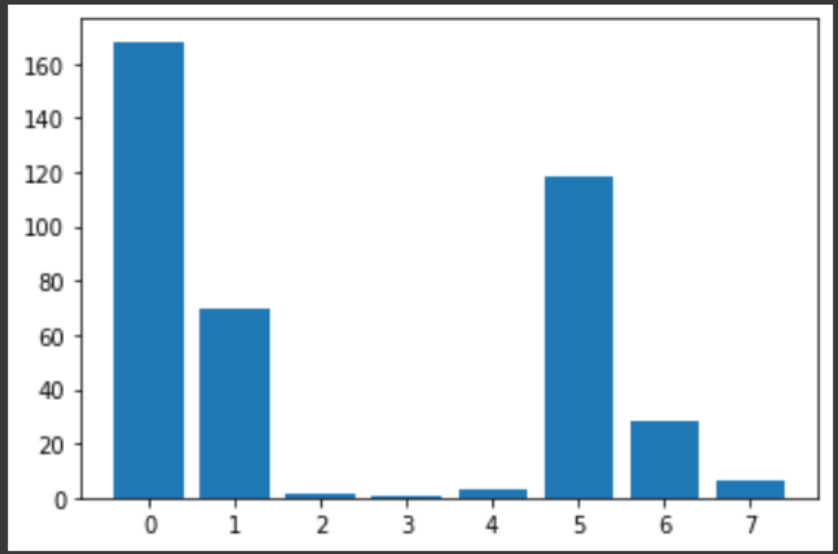


- Berdasarkan Tes Korelasi Numerical Features terhadap Target dengan tes ANOVA, Feature Page Value terlihat berkorelasi kuat dengan Revenue. Feature Exit Rates dan Product Related juga dapat dikatakan cukup berkorelasi dengan Revenue, disusul dengan Product Related Duration dan Bounce Rates.
- Exit Rates dengan Bounce Rates dan Product Related dan Product Related Duration memiliki korelasi yang bersifat redundan atau multicollinearity, sehingga kita memilih Exit Rates dan Product Related untuk diketahui lebih lanjut.

Multivariate Analysis – Categorical Correlation

- Berdasarkan Tes Korelasi Categorical Features terhadap Target dengan tes Chi Square, Feature Special Day terlihat berkorelasi kuat dengan Revenue. Feature Traffic Type dan Month juga dapat dikatakan cukup berkorelasi baik dengan Revenue.
- Korelasi kuat antara Special Day dan Revenue dapat diasumsikan jika mendekati Special Day, user akan melakukan transaksi karena adanya banyak promo diadakan menjelang Special Day atau user ingin memberikan hadiah ke orang lain sehingga perlu membeli sebelum waktu Special Day

```
Feature SpecialDay: 168.130343
Feature Month: 69.565918
Feature OperatingSystems: 1.292720
Feature Browser: 0.456487
Feature Region: 2.910318
Feature TrafficType: 118.427431
Feature VisitorType: 27.954441
Feature Weekend: 6.199053
```



4. Business Insight

Selain EDA, lakukan juga beberapa analisis dan visualisasi untuk menemukan suatu business insight. Tuliskan minimal 3 insight, dan berdasarkan insight tersebut jelaskan rekomendasinya untuk bisnis.

Business Insight & Recommendation

- Jumlah revenue atau pendapatan yang didapat dari pelanggan lama atau yang kembali lebih banyak daripada pelanggan baru. Namun, tingkat konversi pelanggan baru lebih tinggi dibandingkan dengan pelanggan lama. Dari total pengujung sebanyak 85% merupakan pengunjung kembali ke situs dan 15% pengunjung adalah baru. Kita dapat memberikan tawaran atau campaign untuk menarik lebih banyak pengunjung baru agar tertarik melakukan pembelian pada situs web dan membuat pelanggan lama untuk melakukan transaksi kembali di situs web.
- Sebanyak 65% pengunjung berasal dari browser 2 dan lebih dari 85% pengunjung berasal dari browser 1 dan 2. Kita dapat membuat situs web menjadi lebih menarik, interaktif, dan responsif terhadap browser ini. Selain itu, untuk meningkatkan konversi pada browser lainnya, kita dapat memasang iklan situs web pada browser lainnya.
- Wilayah 1 menyumbang penjualan lebih banyak diikuti oleh wilayah 3. Dengan informasi ini, dapat direncanakan campaign dan penyediaan pasokan barang dengan cara yang lebih baik. Sebagai contoh, kita mungkin mengusulkan untuk membangun gudang yang khusus melayani kebutuhan wilayah 1 untuk meningkatkan tingkat pengiriman dan memastikan bahwa produk dengan permintaan tertinggi selalu tersedia dengan baik.
- Pengunjung situs web tertinggi di bulan Mei, tetapi jumlah pembelian atau transaksi paling besar terjadi di bulan November. Hal ini perlu diselidiki lebih lanjut oleh tim bisnis untuk mengetahui apa yang menyebabkan atau faktor yang mempengaruhi tingginya transaksi pada bulan november

Business Insight & Recommendation

- Sekitar 95% pengunjung menggunakan operating system (OS) 1, 2, atau 3. Dengan mengetahui OS apa saja yang sering digunakan pelanggan untuk melakukan transaksi, bisa menjadi bahan pertimbangan jika kita ingin membuat aplikasi belanja yang user friendly. Dengan adanya aplikasi yang tersedia di aplikasi store di masing-masing OS dapat lebih memudahkan customer melakukan pencarian atau pembelian, serta memudahkan kita memberikan promosi dengan membuat notifikasi aplikasi.
- Rata-rata pengeluaran pada halaman administratif, informasi, dan produk terkait lebih tinggi bagi mereka yang membeli sesuatu daripada mereka yang tidak membeli apa-apa.
- Rata-rata Bounce Rate dan Exit Rate lebih rendah saat ada penjualan produk.
- Halaman-halaman dengan Page Values tinggi memiliki bounce rate yang lebih rendah. Kita harus berbicara dengan tim teknologi kami untuk menemukan cara meningkatkan Page Values dari halaman web.
- Konversi pengunjung pada hari weekdays lebih banyak yang tidak melakukan transaksi dibandingkan dengan hari weekend, namun jumlah pengunjung pada hari weekend masih terlalu rendah. Solusi yang akan kami lakukan adalah memprioritaskan pada hari weekend yang memiliki potensi konversi lebih tinggi dari hari weekdays dengan memberikan rekomendasi promosi diskon produk di hari weekend

5. Git

Upload project teman-teman di sebuah repository git. Berkolaborasilah di Git jika ada perubahan version dari waktu ke waktu.

A. Buat Repository Git
B. Upload file notebook atau file pengerjaan lainnya pada repository tersebut

Untuk file README, dapat merupakan summary insight yang telah didapatkan dari EDA.

Link Git ec-Team:
https://github.com/EC-Teams/Final-Project-Online-Shopping-Intention