Stage 1 – Data Wizards

1. Descriptive Statistics (15 poin)

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?

Terdapat beberapa tipe data yang kurang sesuai, seperti Region, TrafficType, Browser, OperatingSystem yang seharusnya menjadi Object.

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

Pada dataset terdapat 12946 baris data, dengan 18 feature. Terdapat data dublikat sebanyak 711 baris data. Feature yang memiliki nilai null yaitu :

- Administrative : 111 nilai kosong
- Administrative_Duration : 633 nilai kosong
- ProductRelated_Duration : 639 nilai kosong
- BounceRates : 74 nilai kosong
- OperatingSystems : 524 nilai kosong

Data pada Feature OperatingSystem, Browser, Region, dan TrafficType akan dilakukan Feature Encoding untuk membantu memahami data tersebut. Berikut adalah pelabelan data :

- **Region :** Jakarta (1), Bandung (2), Surabaya (3), Medan (4), Batam (5), Makassar (6), Tangerang (7), Yogyakarta(8), Semarang (9).
- **Browser :** Safari (1), Google Chrome (2), Internet Explorer (3), Mozilla Firefox (4), Microsoft Edge (5), Samsung Internet (6), Maxthon Browser (7) , Brave (8), Vivaldi (9) UC Browser (10), DuckDuckGo (11), Opera (12), Netscape Navigator (13).
- **OperatingSystem :** iOS (1.0), Windows (2.0), Android (3.0), MAC OS (4.0), Blackberry OS (5.0), Chrome Os (6.0), Unix (7.0), Linux (8.0),

- **Traffic type :** Organic Search (1), Paid Search (2), Direct traffic (3), Social Media(4), Offline Sources (5), Referral Traffic (6), Email Marketing (8), Display Advertising (10), Affiliate Marketing (13),

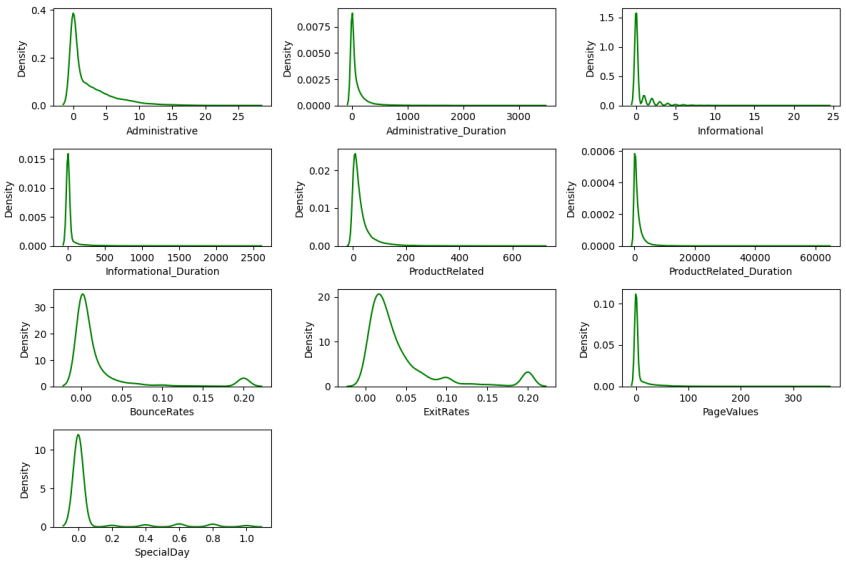
c. Apakah ada kolom yang memiliki nilai summary agak aneh? (min/mean/median/max/unique/top/freq).

Beberapa data memiliki nilai mean yang sangat melebihi nilai mediannya. Hal ini dikarenakan fitur memiliki nilai outlier yang ekstrim.

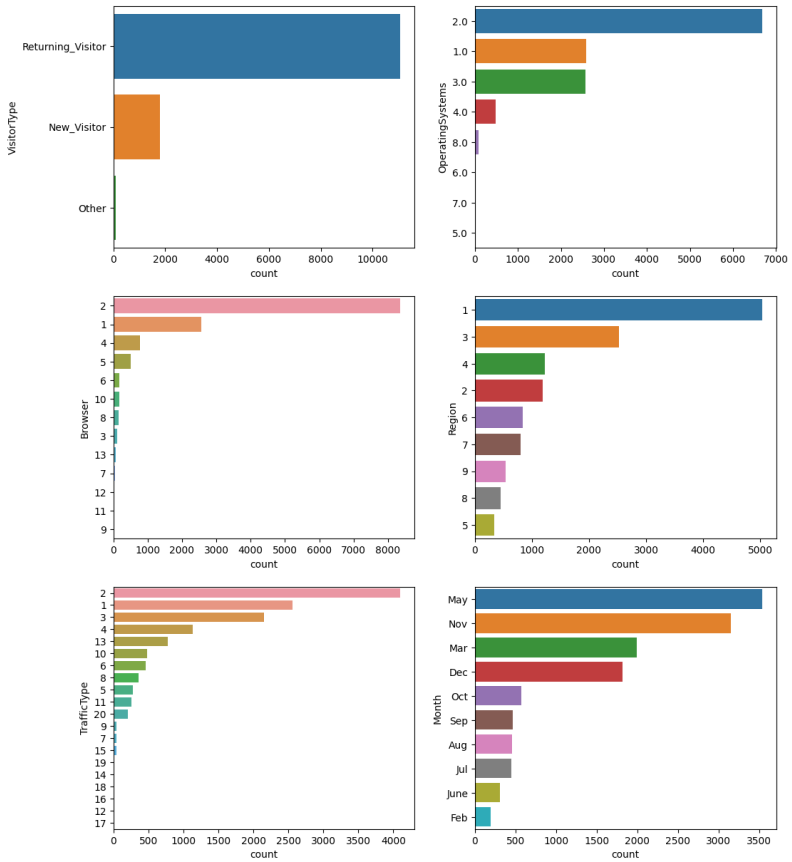
- Pada kolom 'adminstrative' nilai Mean (2.3), Median (1), Q3 (4), dan Max (27).
- Pada kolom 'administrative_duration' Mean (80.37), Median (7), Q3 (92.93) dan Max (3398.75) memiliki perbedaan yang signifikan.
- Pada kolom 'informational' Q3 (0) dan Max (24), Mean (0.4), Median (0) memiliki nilai tidak signifikan.
- Pada kolom 'informational_duration' Mean (34.13), Median (0) dan Q3(0), Max(2549) memiliki nilai berbeda signifikan.
- Pada kolom 'productrelated' Mean (31.65), Median (18) memiliki jarak yang agak jauh dan Q3 (38) dengan Max (705) memiliki nilai yang berbeda signifikan.
- Pada kolom 'productrelated_duration' memiliki nilai Mean (1192.7), Median (599.5), Q3(1470.5), dan Max (63973.5), memiliki perbedaan yang signifikan.
- Pada kolom ‘pagevalues’ Mean (5.8), Median (0), Q3 (0) dan Max (361.763742).
- Pada kolom ‘Month’ tidak terdapat data untuk bulan Januari dan April.

2. Univariate Analysis (25 poin)

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.



Sebagian besar feature memiliki Outlier dan berdistribusi positive skewed, yaitu Administrative, AdministrativeDuration, Informational, InformationalDuration, ProductRelated, ProductRelatedDuration, BounceRates, ExitRates, PageValues, dan Specialday. Untuk menangani data dengan utlier dapat menggunakan Teknik Z-Score atau IQR.



Data kategorik yang memiliki frekuansi tertinggi yaitu :

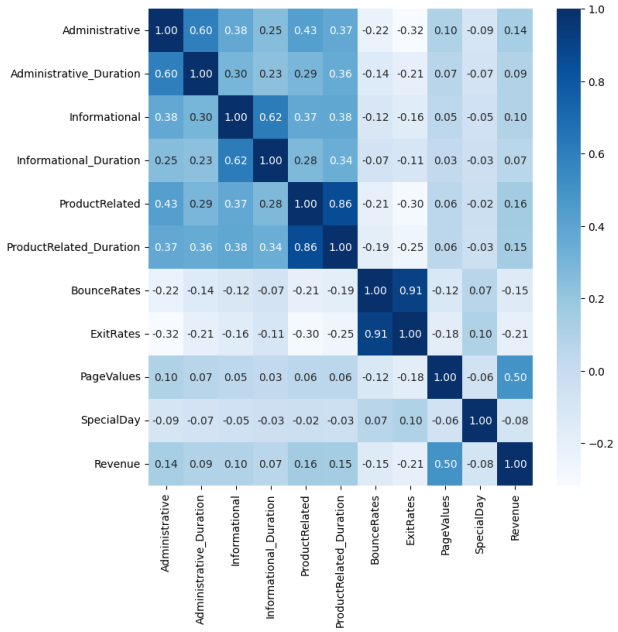
- VisitorType : ReturningVisitor
- OperatingSystem : Windows (2)
- Browser : Google Chrome (2)
- Region : Jakarta (2)
- TrafficType : Paid Search (2)
- Month : May

Data yang memiliki frekuensi yang sangat kecil sehingga tidak terlihat pada grafik yaitu :

- OperatingSystem : Blackberry OS (5), Chrome OS (6), Unix (7).
- Browser : Maxthon (7), Vivaldi (9), DuckduckGo (11), Opera (12), dan Netscape (13).
- TrafficType : 9, 7, 12, 14, 15, 16, 17, 18, 19. Untuk menangani data ini dapat dilakukan drop atau mengabungkannya menjadi ‘Other’.

3. Multivariate Analysis (15 poin)

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?

Berdasarkan korelasi Heatmap tersebut, fitur yang memiliki nilai korelasi yang tinggi yaitu :

- ProductRelated - ProductRelated_Duration sebesar 0.86
- BounceRates - ExitRates sebesar 0.91.

Untuk menangani hal tersebut, akan dilakukan drop pada feature yang memiliki nilai korelasi lebih rendah dengan ‘Revenue’, dengan perbandingan sebagai berukut :

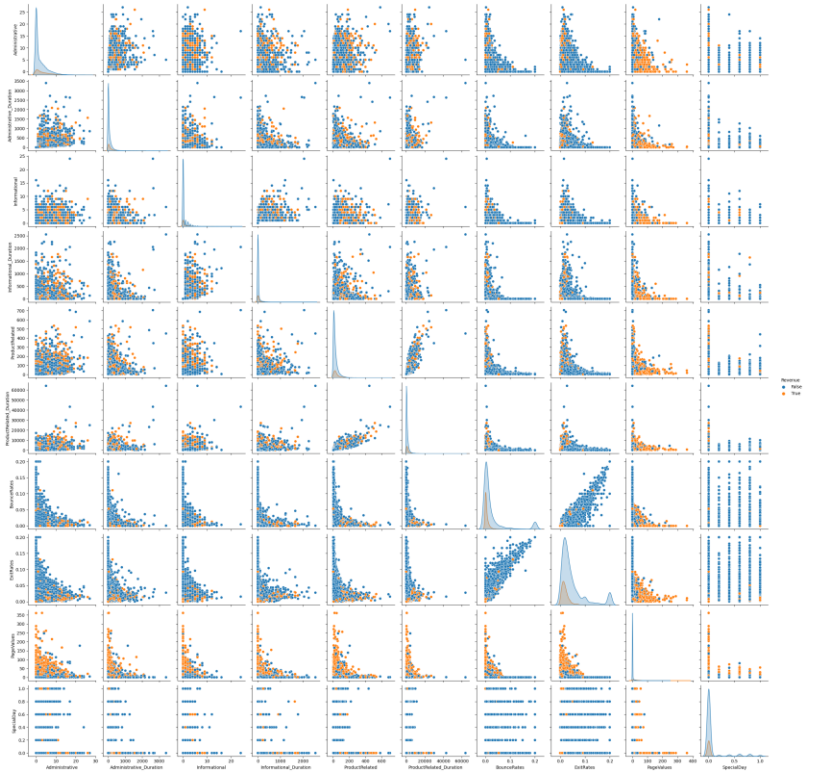
- ProductRelated (0.16) – ProductRelated_Duration (0.15)
- BounceRates (-0.15) - ExitRates (-0.21)

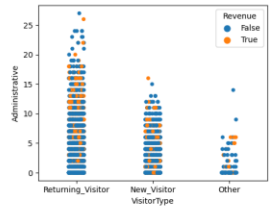
Feature yang memiliki nilai korelasi yang rendah terhadap Revenue yaitu :

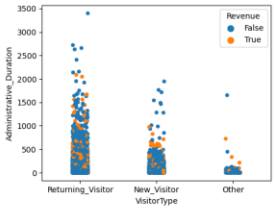
- Administrative_Duration (0.09)
- Informational_Duration (0.07)
- Specialday (-0.08)

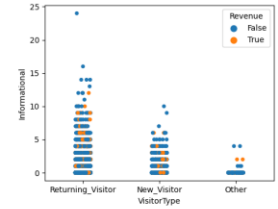
Feature ‘PageValues’ memiliki nilai relasi tertinggi dengan ‘Revenue’.

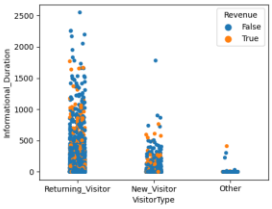
B. Bagaimana korelasi antar-feature, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap feature itu? * Tuliskan juga jika memang tidak ada feature yang saling berkorelasi

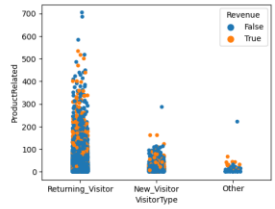


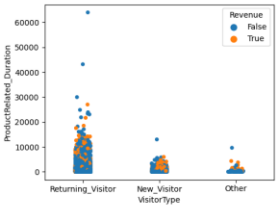

Revenue
False
True
Administrative
Returning_Visitor
New_Visitor
Other
VisitorType

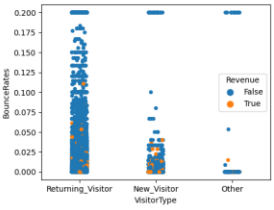

Revenue
False
True
Administrative_Duration
Returning_Visitor
New_Visitor
Other
VisitorType

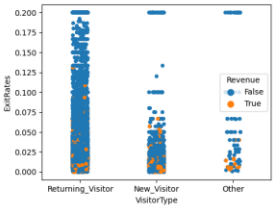

Revenue
False
True
Informational
Returning_Visitor
New_Visitor
Other
VisitorType


Revenue
False
True
Informational_Duration
Returning_Visitor
New_Visitor
Other
VisitorType

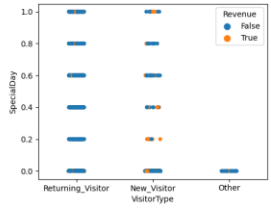

Revenue
False
True
ProductRelated
Returning_Visitor
New_Visitor
Other
VisitorType


Revenue
False
True
ProductRelated_Duration
Returning_Visitor
New_Visitor
Other
VisitorType

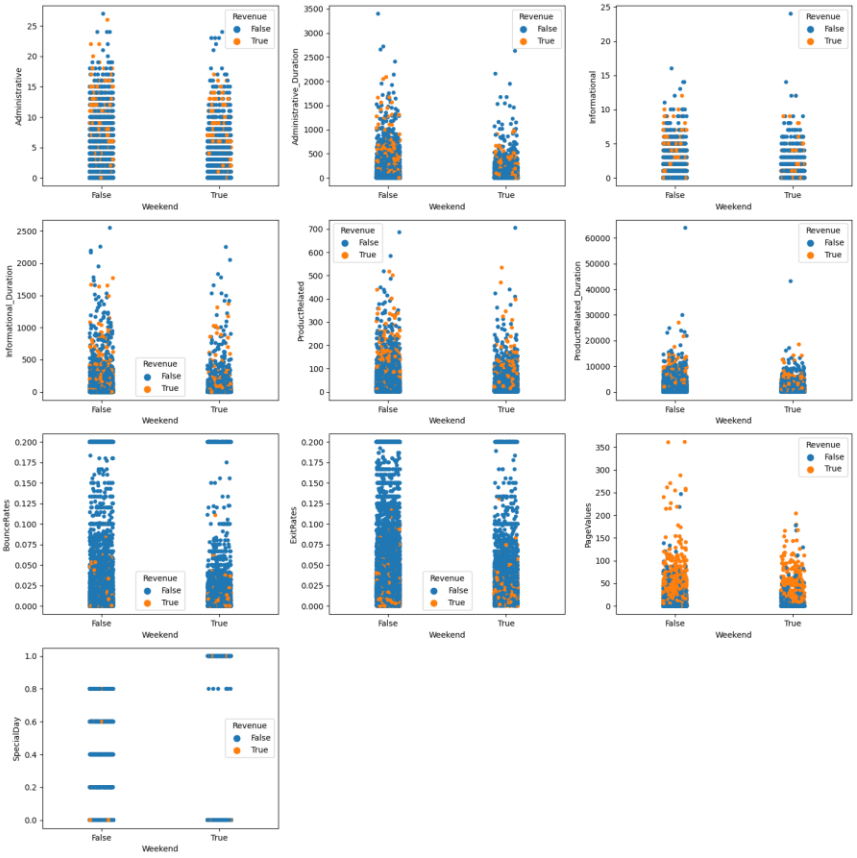

Revenue
False
True
BounceRates
Returning_Visitor
New_Visitor
Other
VisitorType


Revenue
False
True
ExitRates
Returning_Visitor
New_Visitor
Other
VisitorType


Revenue
False
True
PageValues
Returning_Visitor
New_Visitor
Other
VisitorType


Revenue
False
True
SpecialDay
Returning_Visitor
New_Visitor
Other
VisitorType

Pada plot diatas, terlihat bahwa sebaran data BounceRates dan exitrates memiliki korelasi negative, berbanding terbalik dengan PageValues.

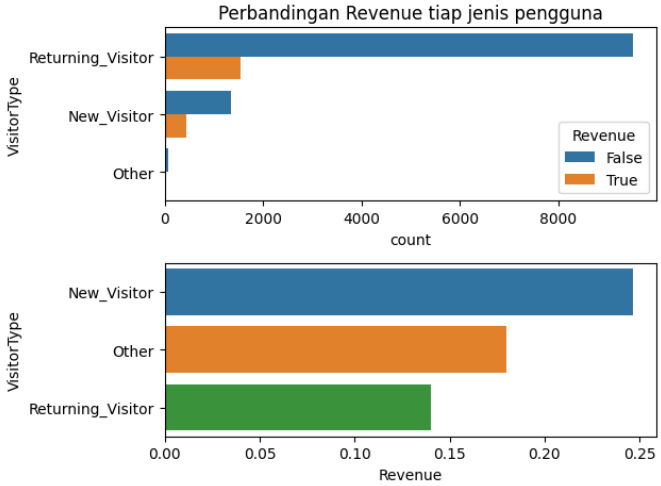


Durasi penggunaan feature pada weekdays memiliki frekuensi yang lebih tinggi dibandingkan dengan weekend.

4. Business Insight (30 poin)

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.

a. Perbandingan Revenue terhadap VisitorType

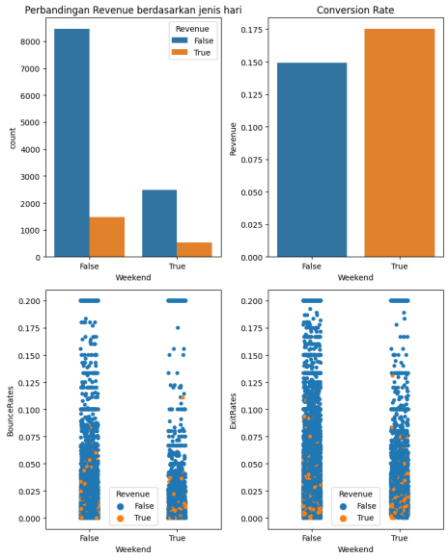


- Nilai Conversion rate tertinggi berasal dari New Visitor (24,64%)
- Kunjungan Tertinggi berasal dari Returning Visitor (11.072 kunjungan)
- New Visitor lebih banyak melakukan pembelian dibandingkan Returning Visitor.

Business Recommendation :

Perusahaan perlu mempelajari hal-hal atau strategi apa saja yang sudah efektif dalam menarik pelanggan baru sebelumnya agar bisa ditingkatkan lagi. Selain itu, perlu dilakukan analisa data perilaku returning visitor untuk mengetahui lebih jauh alasan mereka melakukan pembelian atau tidak. Perusahaan juga bisa menawarkan promo-promo untuk pembeli yang melakukan repeat order atau program loyalty agar bisa meningkatkan minat pembelian pada returning visitor.

b. Perbandingan Revenue terhadap Weekend

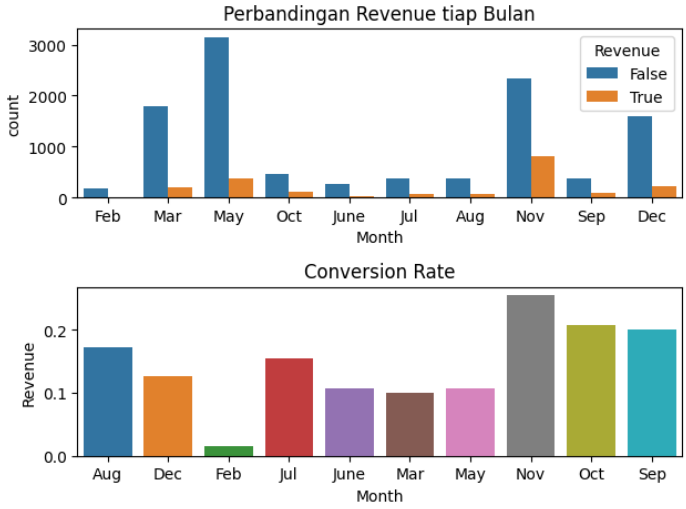


- Nilai Conversion rate tertinggi berasal dari Weekend (17,5%)
- Kunjungan tertinggi berasal dari Weekdays (9929 kunjungan)
- Hal ini dapat disebabkan oleh Tingkat BounceRates dan ExitRates yang tinggi pada Weekdays, dimana BounceRates dan ExitRates memiliki korelasi negative dengan Revenue.

Business Recommendation :

Memberikan promo weekdays yang diberikan pada pelanggan yang melakukan pembelian pada weekend atau memberitahukan promo yang berlaku pada weekdays only untuk pelanggan yang sering berkunjung pada hari weekdays. Optimalkan pengalaman user pada saat weekday untuk mengurangi bounce rates atau exit rates seperti menampilkan halaman-halaman relevan atau promo menarik. Meningkatkan promosi pada saat weekend serta menganalisa strategi apa saja yang telah berhasil meningkatkan conversion rate pada saat weekend. Perusahaan perlu memperluas data mengenai pelanggan agar dapat lebih dalam melakukan segmentasi pelanggan sehingga pemberian promo lebih relevan dan personal, misalnya pada jam berapa pada hari kerja jumlah tertinggi pelanggan melakukan kunjungan, sehingga perusahaan dapat melakukan pemeritahuan promo atau penawaran promo lebih efektif.

c. Perbandingan Revenue terhadap Month

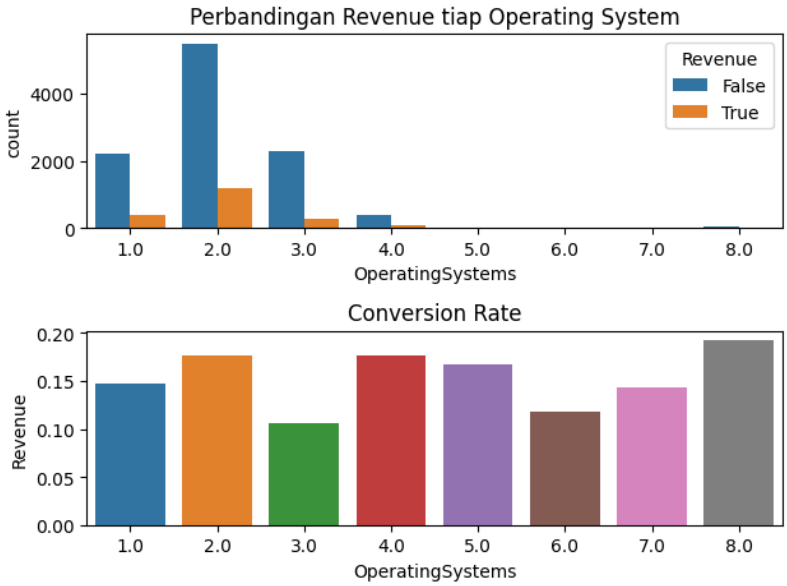


- Nilai Convertion Rate tertinggi terjadi pada bulan November (25%)
- Kunjungan tertinggi terjadi pada bulan Mei (3533 kunjungan)

Business Recommendation :

- Fokuskan upaya pemasaran pada bulan November untuk meningkatkan konversi agar semakin tinggi. Selain itu, dapat pula dilakukan analisa lebih lanjut terhadap hal apa yang mempengaruhi tingkat konversi pada bulan November.
- Optimalkan strategi konversi di Bulan Mei. Meskipun memiliki tingkat kunjungan yang tinggi, bulan Mei menunjukkan tingkat konversi yang relatif lebih rendah dibandingkan dengan bulan November. Untuk mengatasi perbedaan ini, bisnis sebaiknya menganalisis perilaku pengguna dan mengoptimalkan strategi konversi yang khusus ditujukan untuk pengunjung selama bulan Mei.
- Bulan Oktober dan Mei juga perlu dioptimalkan lagi strategi pemasarannya karena jumlah kunjungan yang relatif banyak tapi conversion ratenya masih rendah. dioptimalkan karena memiliki tingkat kunjungan yang relatif lebih tinggi tapi tingkat konversinya.

d. Perbandingan Revenue terhadap OperatingSystem

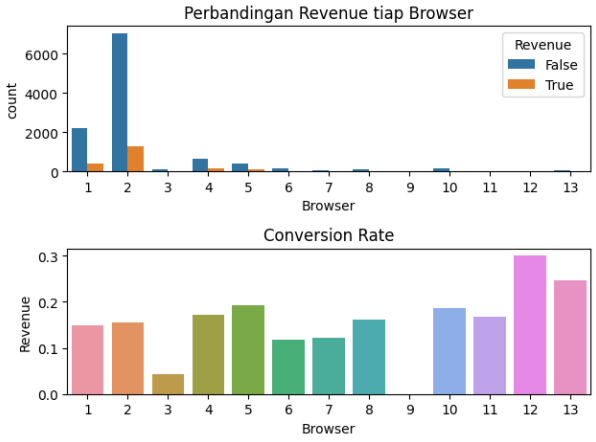


- Nilai Convertion rate tertinggi menggunakan OperatingSystem Linux (8.0)
- OperatingSystem yang paling banyak digunakan yaitu Windows (2.0)
- Nilai Convertion rate dan jumlah kunjungan menggunakan OperatingSystem Linux tidak berbanding lurus sehingga data tidak relevan.

Business Recommendation :

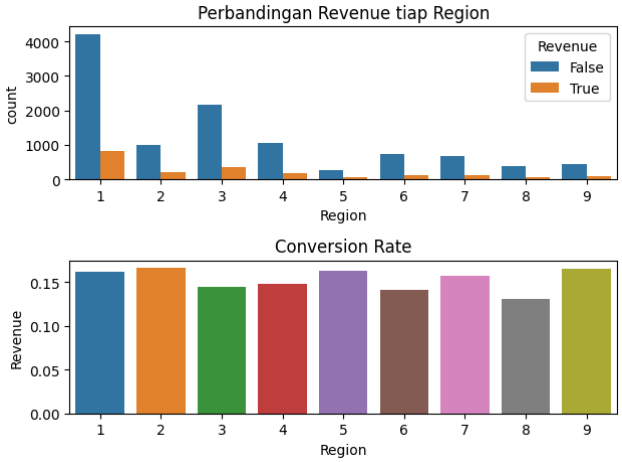
Memberi saran untuk tim developer agar menganalisa serta meningkatkan kompabilitas dan fungsionalitas di berbagai operating system agar kinerja situs web semakin baik.

e. Perbandingan Revenue terhadap Browser



- Nilai Convertion rate tertinggi menggunakan Browser Opera (12)
- Kunjungan tertinggi menggunakan Browser Google Chrome (2)

f. Perbandingan Revenue terhadap Region

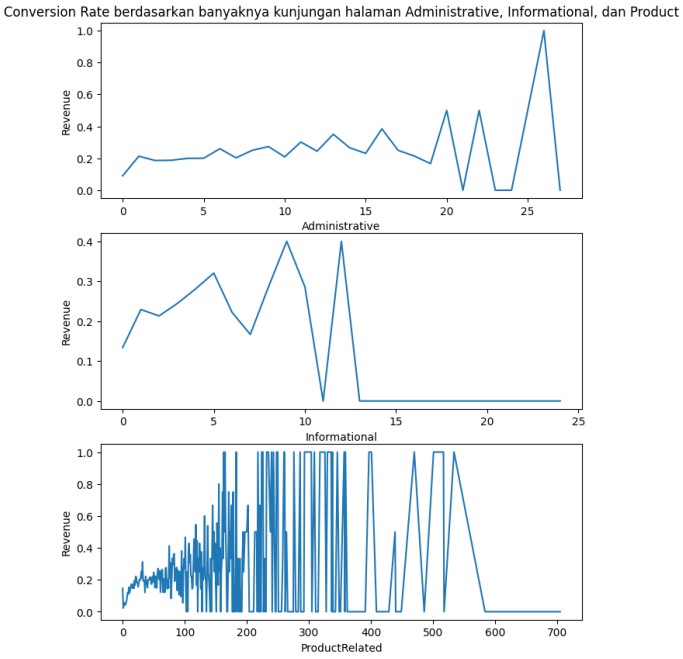


- Nilai Convertion rate tertinggi terdapat pada Region Bandung (2) sebesar 16,63%
- Kunjungan terbanyak terdapat pada Kota Jakarta (1) dan Surabaya (3)

Business Recommendation :

Melakukan analisa lebih lanjut apa yang menyebabkan nilai conversion rate rendah pada beberapa regional. Memahami karakteristik market setiap regional untuk menyesuaikan strategi bisnis yang dapat membantu memaksimalkan konversi setiap regional.

g. Convertion Rate berdasarkan banyaknya kunjungan halaman Administrative, Informational, dan ProductRelated



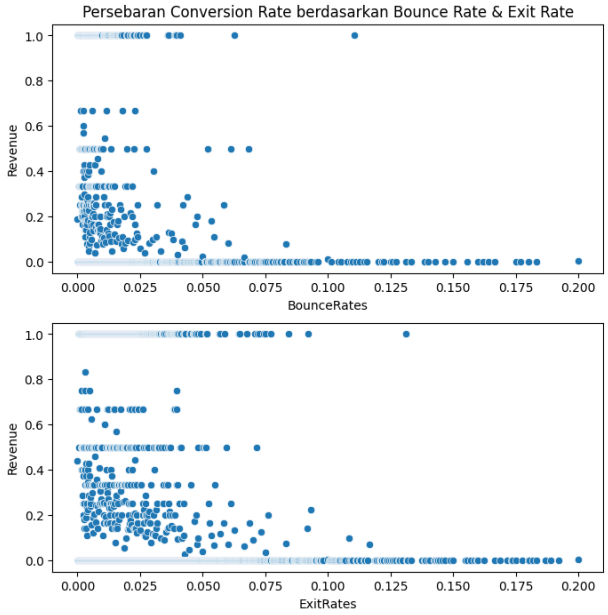
Business insight:

Untuk data administrative, product related dan juga informational terdapat trend positif dimana semakin tinggi nilai pada sumbu X maka semakin tinggi pula tingkat konversinya. Akan tetapi, keberadaan outlier membuat trend grafik menjadi tidak begitu terlihat.

Business Recommendation:

Melakukan analisa lebih lanjut tentang halaman apa saja yang berhasil membuat pelanggan melakukan konversi agar kemudian dapat membuat strategi bisnis dan pemsaran yang lebih efektif.

h. Persebaran Conversion Rate berdasarkan BounceRates dan ExitRates



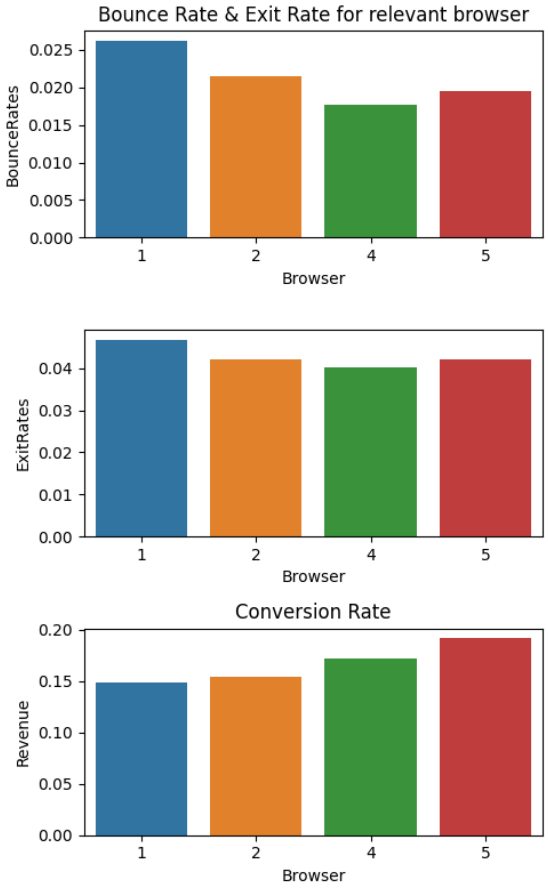
Business Insight:

Semakin tinggi bounce rates dan exit rates, maka tingkat konversi semakin rendah.

Business Recommendation:

Dapat melakukan analisa lebih lanjut tentang alasan yang dapat menyebabkan terjadinya hal tersebut, misalnya foto produk yang kurang menarik, review atau rating pelanggan yang buruk dan sebagainya.

i. BounceRates & Exitrates for relevant Browser



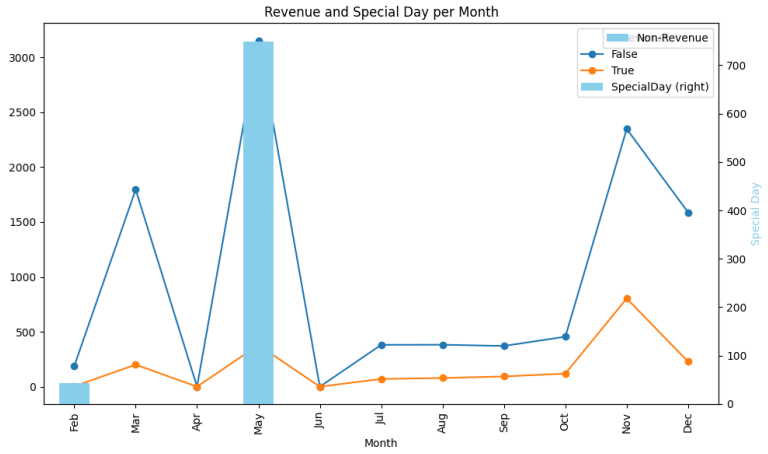
Business insight :

Dapat kita lihat pada browser Safari (1), Google Chrome (2), dan browser Mozilla Firefox (4) semakin tinggi bounce rate dan exit rates maka semakin rendah pula conversion ratenya. Akan tetapi browser Microsoft Edge (5) yang memiliki nilai exit rates dan bounce rates lebih tinggi daripada browser Mozilla Firefox (4) menunjukkan nilai conversion rate yang lebih tinggi.

Business recommendation:

Perlu dilakukan analisa lebih lanjut mengenai hal-hal yang menyebabkan exit dan bounce rates.

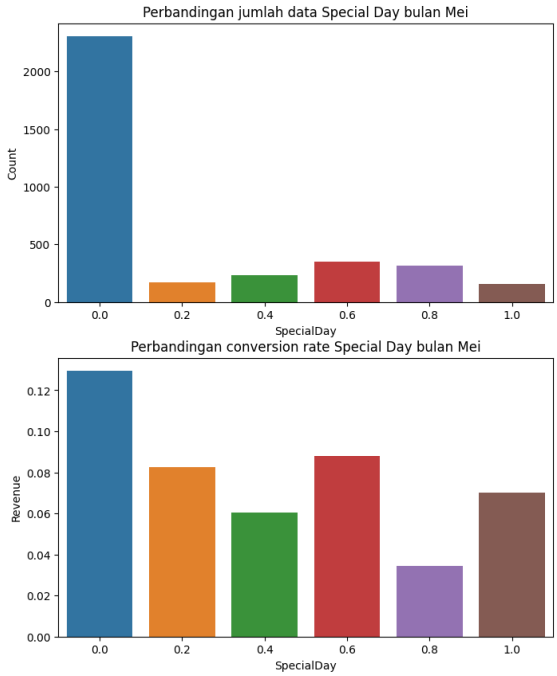
j. Revenue & Special Day by Month



Business Insight :

- Hubungan antara Special Day dan Revenue: Terlihat bahwa bulan-bulan dengan jumlah special day yang lebih tinggi, seperti Mei cenderung memiliki jumlah revenue yang lebih tinggi. Hal ini menunjukkan adanya korelasi antara special day dan aktivitas kunjungan untuk pembelian. Kehadiran special day seperti hari libur atau acara promosi khusus dapat memengaruhi tingkat penjualan.
- Polanya Tidak Konsisten: Walaupun terdapat bulan-bulan dengan jumlah revenue yang tinggi, namun terdapat bulan-bulan lain yang memiliki jumlah revenue yang lebih rendah atau bahkan nol. Hal ini menunjukkan adanya fluktuasi dalam performa penjualan dari bulan ke bulan, yang mungkin dipengaruhi oleh banyak faktor seperti tren pasar, perubahan perilaku konsumen, atau strategi pemasaran yang berbeda.
- Pada bulan Mei, hampir semua kunjungan

k. Perbandingan Jumlah data dan Conversion Rate Special Day pada bulan Mei



Business Insight :

- Pada bulan Mei, pengguna yang beraktivitas mendekati special day cukup banyak. Hal ini mungkin yang memengaruhi banyaknya jumlah pengguna saat bulan Mei. Namun, conversion rate saat mendekati special day masi lebih rendah dibanding hari biasa. Kemungkinan pengguna hanya melihat-lihat barang tanpa melakukan pembelian.
- Selain itu, pada saat menjelang special day terdapat peningkatan conversion rate pada skala 0.6

Business Recommendation :

- Pada bulan Mei, adakan banyak promo untuk memaksimalkan banyaknya pengguna.
- Melakukan analisa lebih lanjut terhadap faktor yang mempengaruhi peningkatan conversion rate pada skala 0.6 agar dapat mengoptimalkan strategi pemasaran selanjutnya seperti menambahkan promo menjelang hari spesial, pemasangan iklan atau rekomendasi produk sesuai hari spesial dan lain- lain.