Rakamin Academy

Homework EDA

Kelompok 2

Agustina Sri Wardani
Fatchul Arifin
Ferry Setefanus
Gigih Septian
Kornelius Rio
M. Harun Arrasyid
Raza Aqil Maulana

Final Project - Stage 1

1. Descriptive Statistics

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

- kolom 'children': lebih tepat bertipe int64 bukan float64 karena tidak ada hitungan anak yang berkoma seperti 2.5
- kolom 'agent'dan 'company': kedua kolom ini berisi data ID.
  Idealnya, tipe data untuk ID adalah int64 atau object. Untuk kolom 'agent'dan 'company' lebih tepat menggunakan int64 karena dilihat dari dataset ID kedua kolom ini dalam angka
- kolom reservation_status_date: harusnya bertipe datetime bukan object

Coding selengkapnya dapat dilihat disini

Disini juga bisa, yuk mampir

1. Descriptive Statistics

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

Kolom yang memiliki nilai kosong

- 'company' sebesar 94.307%.

  Kolom 'company' akan di drop karena nilai kosong terlalu besar
- 'agent' sebesar 13.686%.

  Kolom 'agent' akan diubah dengan value 0: tidak punya ID dan 1: punya ID
- 'country' sebesar 0.409%.

  Nilai kosong akan diisi dengan mode (country yang paling sering muncul)
- 'children' sebesar 0.003%.

  Nilai kosong akan di drop

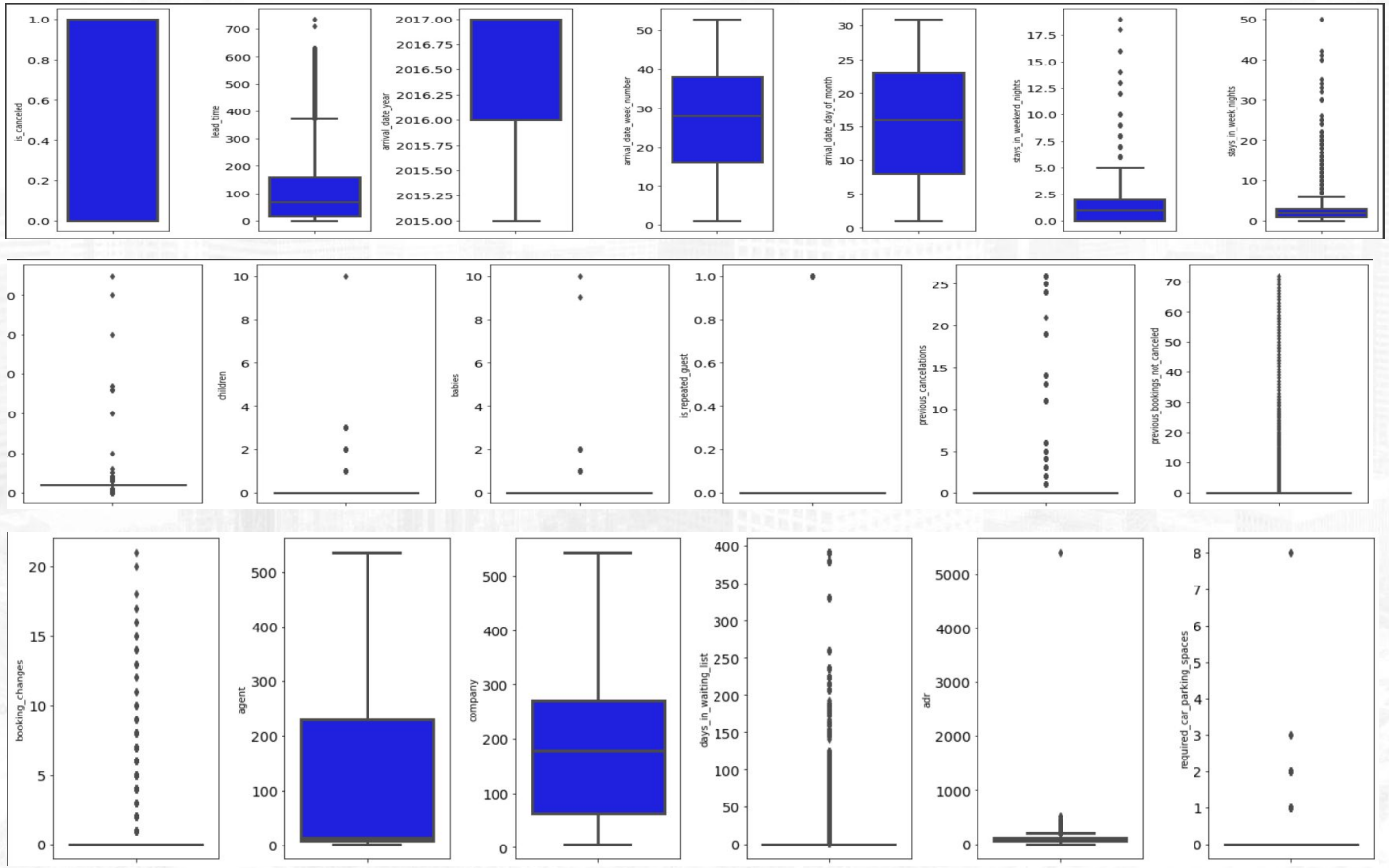
1. Descriptive Statistics

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

- Pada kolom **lead_time** terdapat nilai **max** yang cukup aneh yaitu sebesar 737 yang artinya jarak antara hari dilakukan reservasi dengan hari-H nya berjarak 737 hari (2 tahun lebih)
- Nilai **max** pada kolom **days_in_waiting_list** adalah 391 yang berarti lamanya konsumen berada pada waiting list untuk mengonfirmasi reservasi yang dilakukan oleh konsumen yaitu 391 hari
- Pada kolom **ADR** terdapat **nilai minus**. ADR (Average Daily Rate) dihitung dari pendapatan rata-rata yang diperoleh dari kamar dan membaginya dengan jumlah kamar yang terjual. Sehingga tidak normal ADR bernilai minus
- Pada kolom **adult** terdapat **nilai 0** (nol) sehingga perlu dilihat lagi mengingat sepertinya tidak mungkin membuat pesanan tanpa adult
- Kolom **country** memiliki terlalu **banyak unique value**
- Kolom **name, email, phone_number, credit_card** memiliki **unique value yang terlalu banyak dan merupakan data pribadi**
- Banyak **data skewed** (mean <> median) yg kemungkinan terdapat **outliers**

Handling penemuan nilai summary agak aneh akan dilakukan pada tahap Data Pre-Processing

2. Univariate Analysis

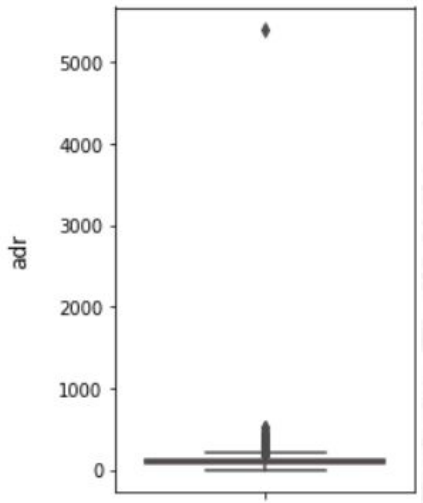


2. Univariate Analysis

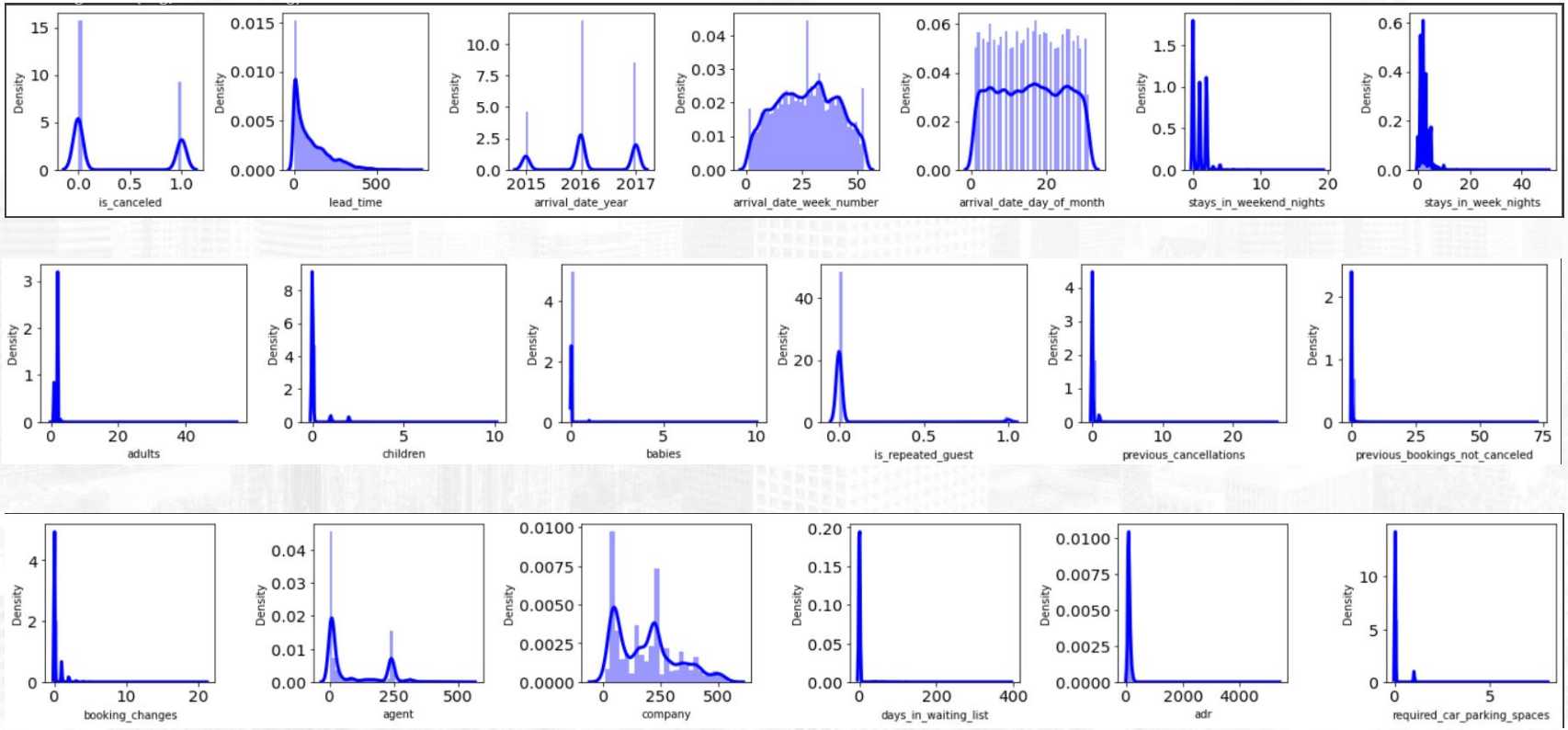
Beberapa informasi penting yang kami amati:

- Mayoritas kolom nums masih memiliki outlier. Tapi sepertinya nilai outlier ini dapat memberikan banyak informasi berguna, jadi kami tidak melakukan data cleaning outlier, melainkan feature engineering atau membuat kolom baru kemudian menggunakan metode z-score.
- Outlier yang akan dihapus adalah yang ada di kolom ADR karena tampaknya memiliki satu nilai yang sangat berbeda
- Dari visualisasi boxplot di atas terlihat bahwa sebagian besar kolom memiliki distribusi yang miring (skewed distribution), kecuali kolom arrival_date_day_of_month, arrival_date_week_number

Dari visualisasi boxplot di samping dapat diketahui kolom 'adr' memiliki satu outlier yang cukup ekstrim bernilai 5400. Data 'adr' yang lain berada di bawah 1000



2. Univariate Analysis

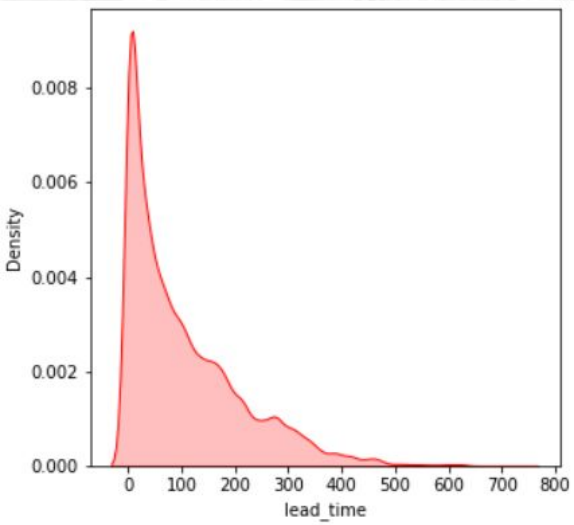


2. Univariate Analysis

Hal terpenting yang harus diperhatikan dari visualisasi displot di atas adalah bentuk distribusinya

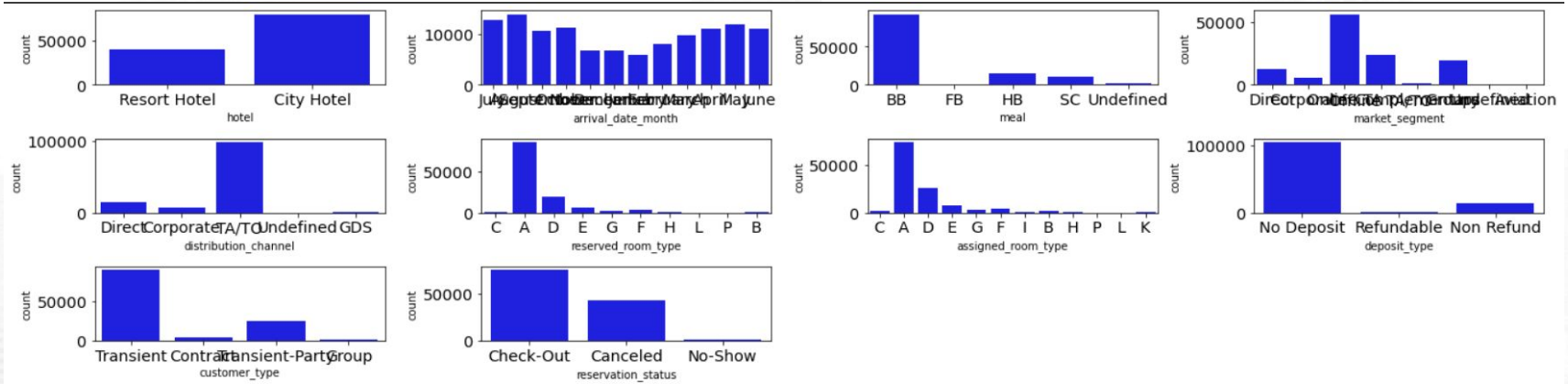
Seperti yang kita duga di visualisasi boxplot, sebagian besar kolom memiliki distribusi miring (skewed distribution), yaitu lead time, stays_weekend_nights, stay_in_week_nights, adults, chidlren, babies, is_repeat_guest, previous_cancellations, previous_booking_not_cancelled, booking_changes, days_in_waiting_list, adr, dan required_car_parking_spaces

Dari visualisasi displot di atas diketahui kolom-kolom skewed ini adalah right skewed sehingga akan dilakukan log transformation pada tahap data pre-processing



Dari visualisasi kdeplot di samping dapat kita lihat bahwa kolom 'lead_time' memiliki distribusi yang positive skewed. Dari visualisasi juga dapat kita lihat terjadi sedikit lonjakan pada lead_time mendekati 300 hari. 'lead_time' juga terus memiliki ekor hingga melebihi angka 700 yang berarti ada customer yang memesan untuk lebih dari 700 hari yang akan datang.

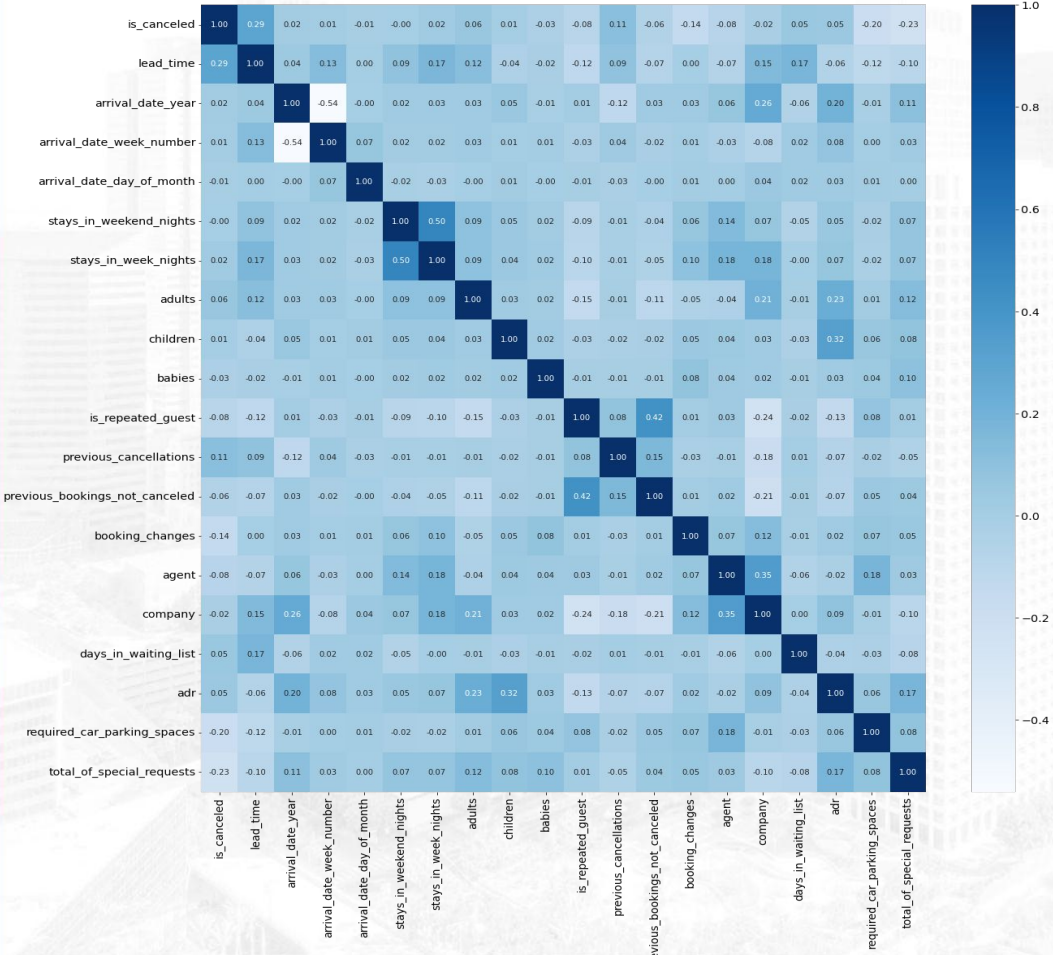
2. Univariate Analysis

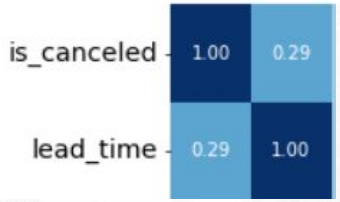


Dari visualisasi countplot di atas diketahui bahwa:

- Pada kolom deposit_type, no deposit menjadi nilai yang mendominasi.
- Kolom name, email, phone-number, credit_card, country, reservation_status_date memiliki kategori (unique value) yang terlalu banyak sehingga memiliki potensi untuk tidak dijadikan sebagai feature untuk model machine learning.

3. Multivariate Analysis



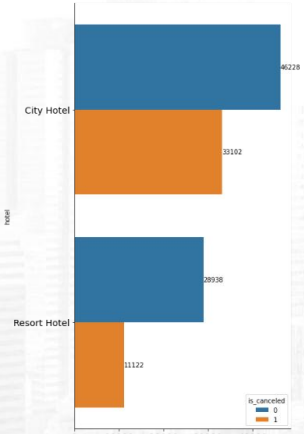


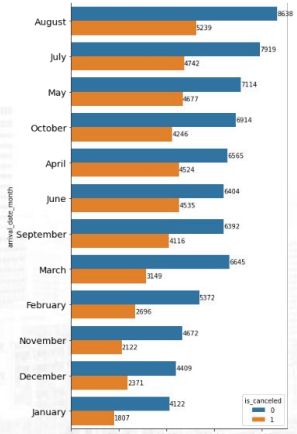
Dari visualisasi correlation heatmap di atas dapat kita ketahui bahwa:

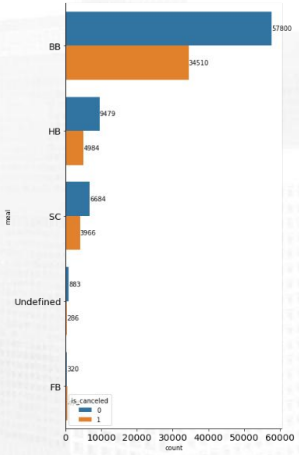
- Target kita (is_cancelled) memiliki korelasi tertinggi yang positif yaitu dengan feature lead_time yaitu 0.29
- Feature lain yang juga memiliki korelasi positif yang cukup besar dengan target adalah previous_cancellations sebesar 0.11

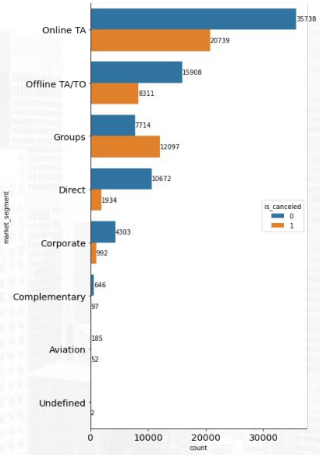
3. Multivariate Analysis

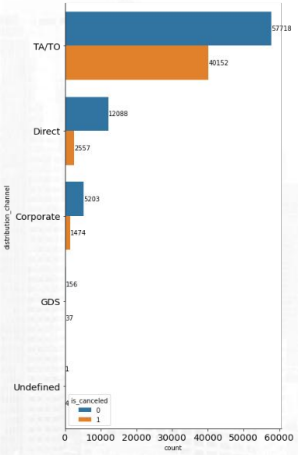
Rakamin Academy

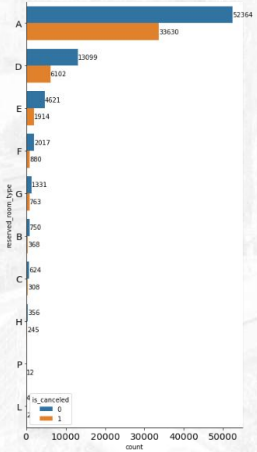


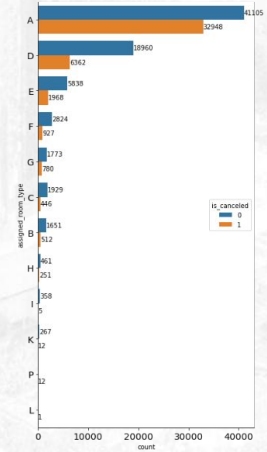


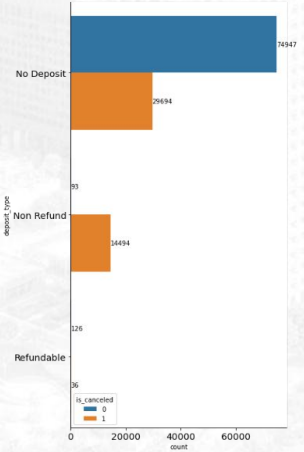


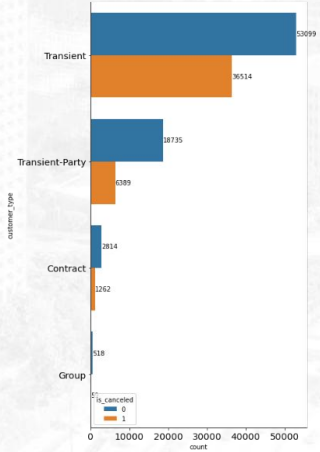


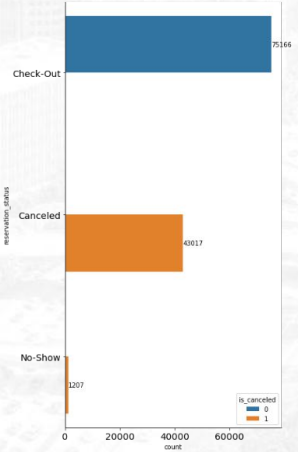












3. Multivariate Analysis

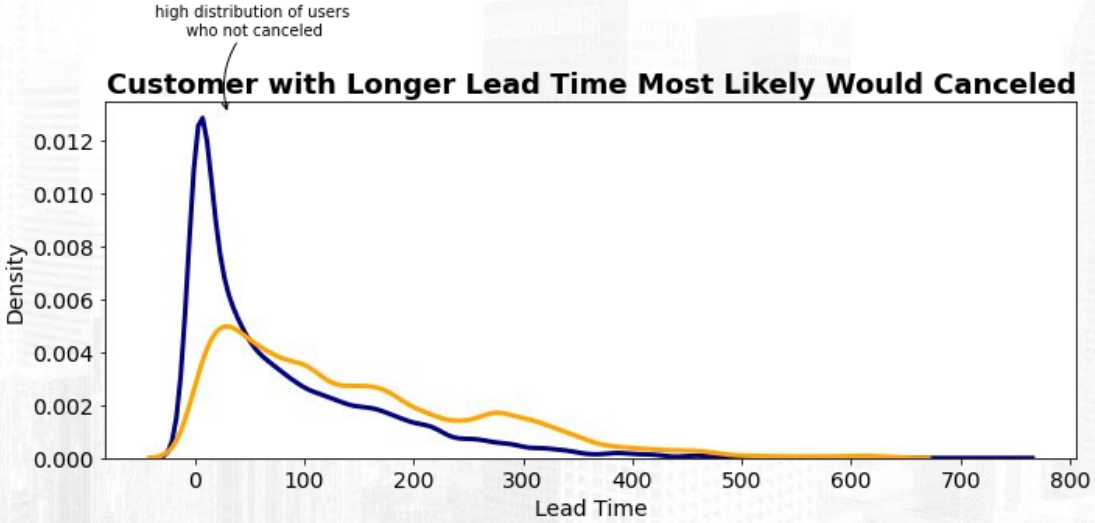
- Pada kolom city hotel, terlihat bahwa jenis city hotel merupakan jenis yang paling banyak dipesan oleh pelanggan dan juga paling banyak dicancel
- Pada kolom arrival_date_month, bulan agustus menjadi bulan yang paling banyak dipilih pelanggan dan paling banyak terjadi cancel
- Pada kolom meal, secara keseluruhan jenis BB paling banyak dipilih dibandingkan dengan jenis lainnya, sehingga cancel juga banyak terjadi pada pelanggan yang memilih jenis BB
- Pada kolom market segment dan distribution channel didominasi oleh customer yang menggunakan travel agent/tour operator untuk melakukan booking. Hal ini membuktikan bahwa customer lebih menyukai reservasi dengan menggunakan platform travel agengt/tour operator
- Pada kolom reserved_room_type dan assigned_room_type menunjukkan bahwa customer lebih menyukai memesan room type A dibandingkan dengan room lainnya.
- Pada kolom deposit type didominasi oleh customer yang melakukan reservasi tanpa deposit.
- Pada kolom customer type, didominasi oleh tipe customer transient yaitu tipe customer yang melakukan reserverasi untuk waktu yang singkat

3. Multivariate Analysis

Pengujian dengan chi square dilakukan untuk melihat korelasi antara kolom categorical dengan kolom target

- Korelasi antara feature dan label dapat dilihat dari plot heatmap correlation di atas, beberapa kolom yang memiliki korelasi di bawah 0.05 atau -0.05 dengan label tidak akan terlalu dieksplor pada tahap data preprocessing. Sehingga **fitur yang relevan dan harus dipertahankan untuk dieksplor** yaitu fitur yang memiliki korelasi lebih dari 0.05 atau -0.05 pada **kolom numerik** yaitu lead time, adults, is_repeated_guest, previous_cancellations, previous_booking_not_canceled, booking_changes, agent, days_in_waiting_list, adr, required_car_parking_spaces, total_of_special_requests, sedangkan untuk **kolom categorical** menurut uji chi square, semua kolom memiliki potensi untuk dieksplor.
- Ditemukan adanya pola yang menarik pada korelasi antar feature yaitu pada feature stay_in_weekend_nights dan stays_in_week_nights serta kedua feature tersebut memiliki korelasi bernilai absolut yang cukup tinggi yaitu 0.5, ada kemungkinan redundan, sehingga akan dilakukan penggabungan kedua feature tersebut menjadi weekend_or_weekdays (berisikan 1 untuk weekend, 0 untuk weekdays). Sepertinya reserved_room_types dan assigned_room_types juga bisa digabungkan menjadi assigned_vs_reserved (1 untuk reserved == assigned, 0 untuk reserved != assigned)

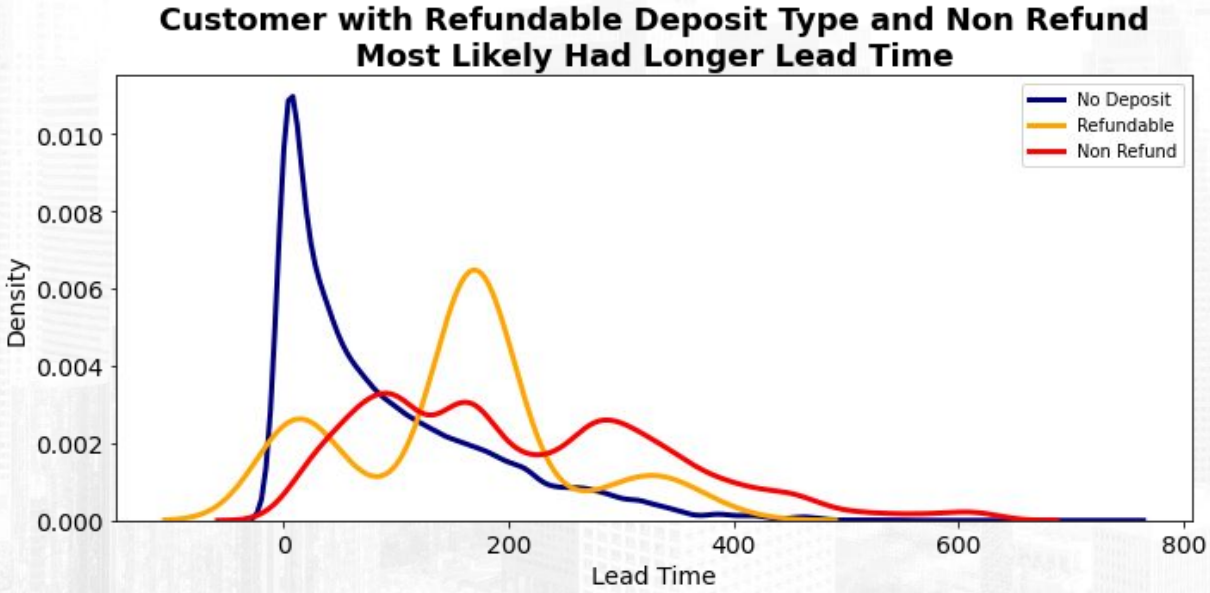
4. Business Insight



Customer yang memiliki lead time yang lebih lama memiliki tingkat canceled yang lebih tinggi.

Rekomendasi bisnis untuk mengatasi masalah tersebut adalah dengan membuat regulasi minimal open reserved seperti 3 bulan - 6 bulan sebelum hari - H. Hal ini juga akan memudahkan pihak hotel untuk menerapkan pricing room secara dinamis tergantung event/season yang akan terjadi pada tanggal yang di booking oleh konsumen.

4. Business Insight

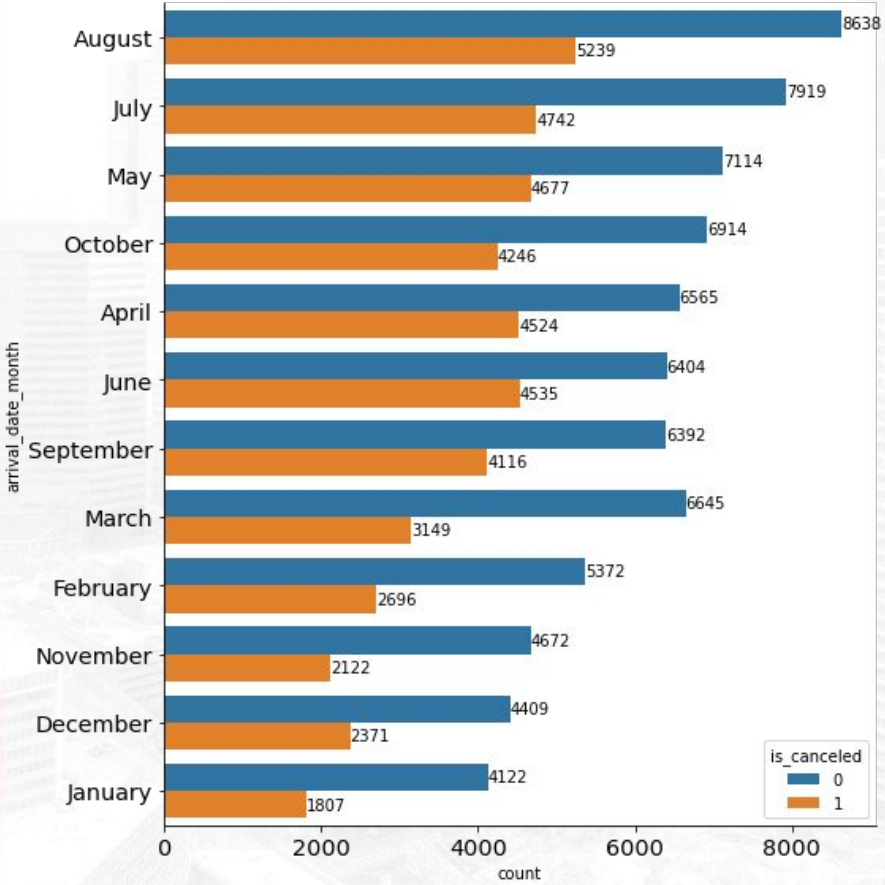


Pemesanan hotel dengan jenis deposit non refund memiliki cancellation rate terbesar yaitu mencapai 99% dibandingkan dengan jenis deposit lainnya.

Salah satu alasannya adalah karena pemesanan hotel dengan jenis deposit non refund memiliki median lead time tertinggi.

Sejalan dengan **rekomendasi bisnis pertama** untuk mengatasi masalah ini adalah dengan membuat regulasi minimal open reserved seperti 3 bulan - 6 bulan sebelum hari - H, lebih dari itu open reserved tidak diterima. Jenis deposit non refund akan diterapkan pada pemesanan dengan lead time lebih dari 6 bulan.

4. Business Insight



Pemesanan Hotel

Apabila dilihat dari waktu kedatangan *customer* setiap bulannya, terlihat bahwa mereka paling banyak datang di musim panas (May, July, Agustus). Kemudian di bulan Agustus terbanyak dikarenakan pada bulan tersebut merupakan puncak musim panas. Rekomendasi bisnis untuk insight ini adalah melalukan summer sale untuk menjangkau lebih banyak konsumen serta membuat tingkat cancel lebih rendah.

Terima Kasih