

## Homework Final Project stage 1

### Kelompok:

1. Abdussalam Darmaatmaja
2. Dimas Jabbar
3. Bijak Ika Handhika
4. Rahmatian Jayanty Sholichah
5. Trully Ananda
6. Nathanael

Data: <https://www.kaggle.com/datasets/mojtaba142/hotel-booking>

### STAGE 1 FINAL PROJECT

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

#### Jawab:

Pertama, import dataset dengan pandas library  
df=pd.read\_csv('hotel\_bookings\_data.csv')

kemudian memeriksa tipe data dan nilai null dari kumpulan data menggunakan info  
df.info()

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119390 entries, 0 to 119389
Data columns (total 36 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   hotel                                119390 non-null  object
1   is_cancelled                         119390 non-null  int64
2   lead_time                           119390 non-null  int64
3   arrival_date_year                   119390 non-null  int64
4   arrival_date_month                  119390 non-null  object
5   arrival_date_week_number            119390 non-null  int64
6   arrival_date_day_of_month           119390 non-null  int64
7   stays_in_weekend_nights             119390 non-null  int64
8   stays_in_week_nights                119390 non-null  int64
9   adults                              119390 non-null  int64
10  children                            119386 non-null  float64
11  babies                             119390 non-null  int64
12  meal                                119390 non-null  object
13  country                             118902 non-null  object
14  market_segment                      119390 non-null  object
15  distribution_channel                119390 non-null  object
16  is_repeated_guest                   119390 non-null  int64
17  previous_cancellations               119390 non-null  int64
18  previous_bookings_not_cancelled     119390 non-null  int64
19  reserved_room_type                  119390 non-null  object
20  assigned_room_type                  119390 non-null  object
21  booking_changes                     119390 non-null  int64
22  deposit_type                        119390 non-null  object
23  agent                               103050 non-null  float64
24  company                             6797 non-null   float64
25  days_in_waiting_list                119390 non-null  int64
26  customer_type                       119390 non-null  object
27  adr                                  119390 non-null  float64
```

Ada beberapa kesalahan di antaranya adalah:

- Kolom "is\_cancelled", "is\_repeated\_guest", "previous\_cancellations", "previous\_bookings\_not\_cancelled" lebih tepat jika tipe datanya boolean. Namun dataset, mempunyai tipe data integer

1	is_canceled	119390	non-null	int64
16	is_repeated_guest	119390	non-null	int64
17	previous_cancellations	118565	non-null	int64
18	previous_bookings_not_canceled	118565	non-null	int64

- Kolom "children","agent","company" akan lebih tepat menggunakan tipe data integer, dikarenakan kolom itu menunjukkan jumlah dengan nilai bulat, sehingga perlu diubah data type float menjadi int64

10	children	119386	non-null	float64
23	agent	103050	non-null	float64
24	company	6797	non-null	float64

- Kolom "reservation\_status\_date" mempunyai data type berupa object, namun lebih tepat bila menggunakan tipe data datetime

31	reservation_status_date	119390	non-null	object
----	-------------------------	--------	----------	--------

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

Jawab:

kita dapat melihat bahwa kolom 'children', 'country', 'agent' dan 'company' memiliki nilai null

stays_in_week_nights	0
adults	0
children	4
babies	0
meal	0
country	488
market_segment	0
distribution_channel	0
is_repeated_guest	0
previous_cancellations	0
previous_bookings_not_canceled	0
reserved_room_type	0
assigned_room_type	0
booking_changes	0
deposit_type	0
agent	16340
company	112593
days_in_waiting_list	0
customer_type	0
adr	0

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

Jawab:

Dari hasil pengecekan tidak ada kolom yang memiliki nilai data yang aneh

```
[28]: df.describe()
```

	is_canceled	lead_time	arrival_date_year	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights	adults	chi
count	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.0
mean	0.370416	104.011416	2016.156554	27.165173	15.798241	0.927599	2.500302	1.856403	0.1
std	0.482918	106.863097	0.707476	13.605138	8.780829	0.998613	1.908286	0.579261	0.3
min	0.000000	0.000000	2015.000000	1.000000	1.000000	0.000000	0.000000	0.000000	0.0
25%	0.000000	18.000000	2016.000000	16.000000	8.000000	0.000000	1.000000	2.000000	0.0
50%	0.000000	69.000000	2016.000000	28.000000	16.000000	1.000000	2.000000	2.000000	0.0
75%	1.000000	160.000000	2017.000000	38.000000	23.000000	2.000000	3.000000	2.000000	0.0
max	1.000000	737.000000	2017.000000	53.000000	31.000000	19.000000	50.000000	55.000000	10.0

children	babies	is_repeated_guest	previous_cancellations	previous_bookings_not_canceled	booking_changes	agent	company	days_in_waiting_list
19390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000	119390.000000
0.103886	0.007949	0.031912	0.087118	0.137097	0.221124	74.828319	10.775157	2.321149
0.398555	0.097436	0.175767	0.844336	1.497437	0.652306	107.141953	53.943884	17.594721
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	7.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	9.000000	0.000000	0.000000
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	152.000000	0.000000	0.000000
10.000000	10.000000	1.000000	26.000000	72.000000	21.000000	535.000000	543.000000	391.000000

adr	required_car_parking_spaces	total_of_special_requests	total_guests	stay_duration
119390.000000	119390.000000	119390.000000	119390.000000	119390.000000
101.831122	0.062518	0.571363	1.968239	3.427900
50.535790	0.245291	0.792798	0.722394	2.557439
-6.380000	0.000000	0.000000	0.000000	0.000000
69.290000	0.000000	0.000000	2.000000	2.000000
94.575000	0.000000	0.000000	2.000000	3.000000
126.000000	0.000000	1.000000	2.000000	4.000000
5400.000000	8.000000	5.000000	55.000000	69.000000

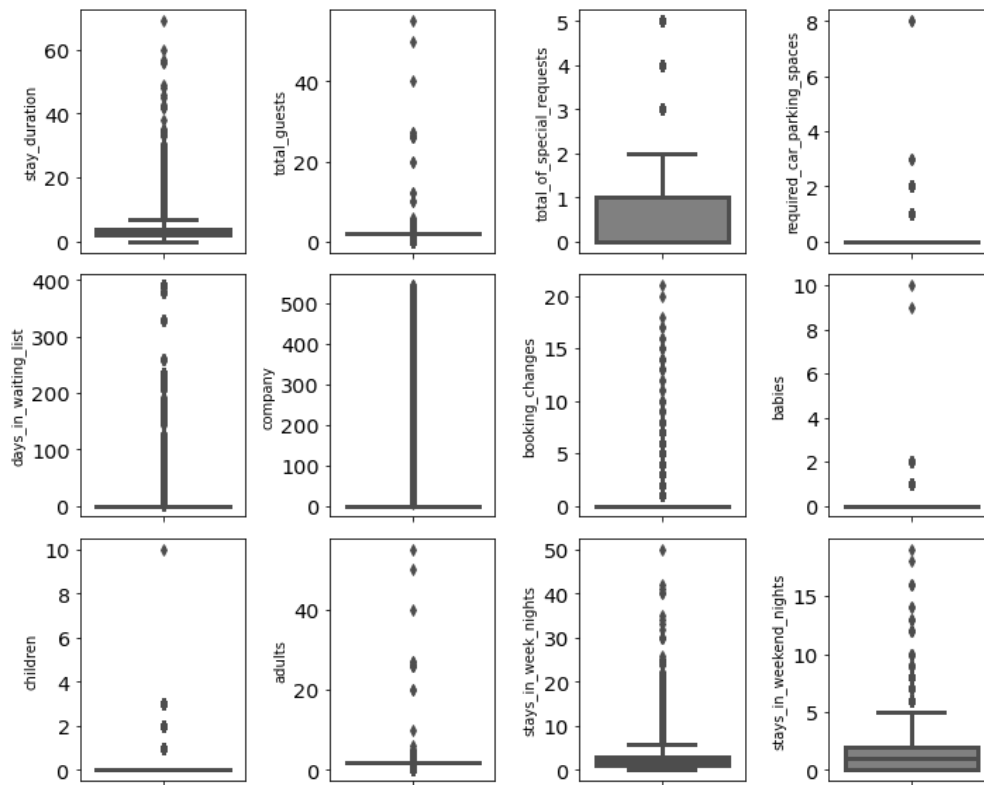
## 2. Univariate Analysis

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.

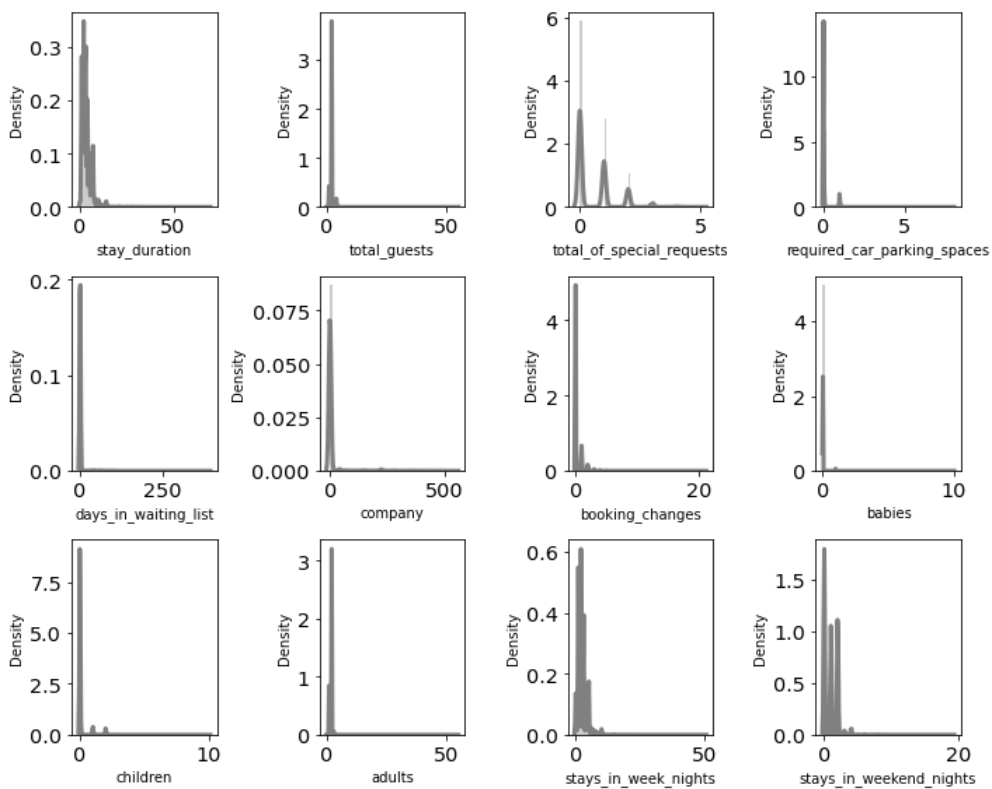
Jawab:

Kolom yang divisualisasikan adalah kolom yang mengandung data numerik.

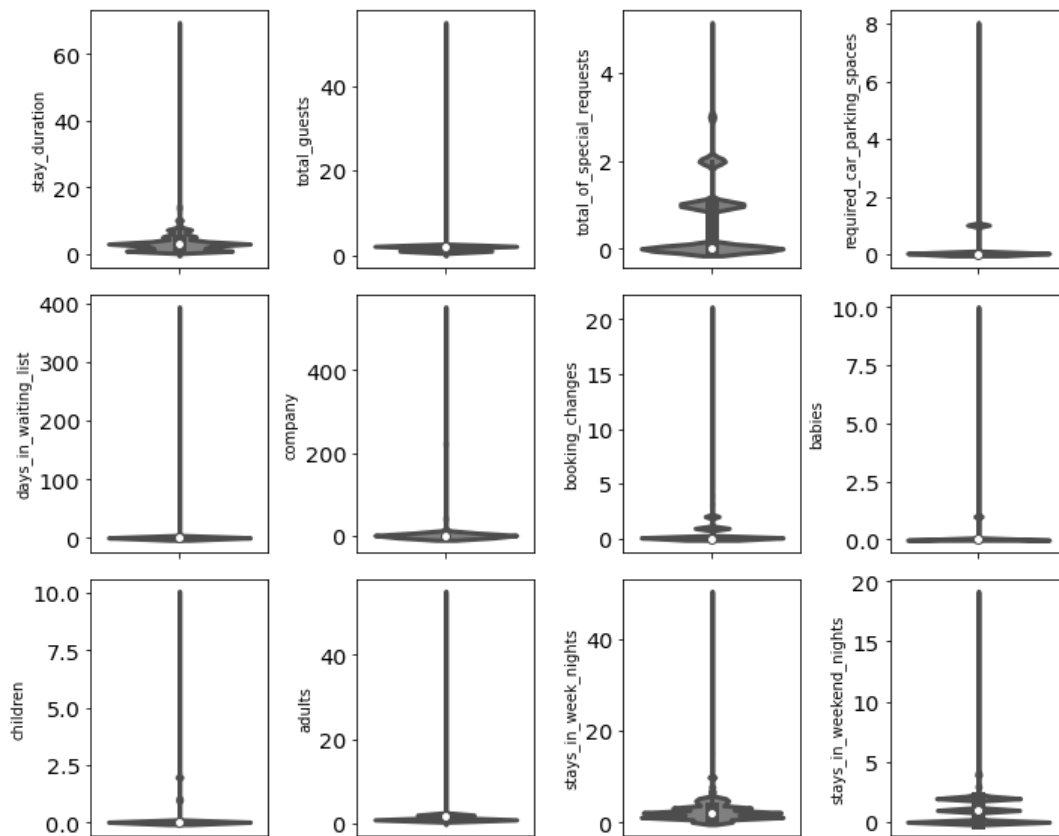
a. Box Plot



b. Distribution Plot



### c. Violin plot

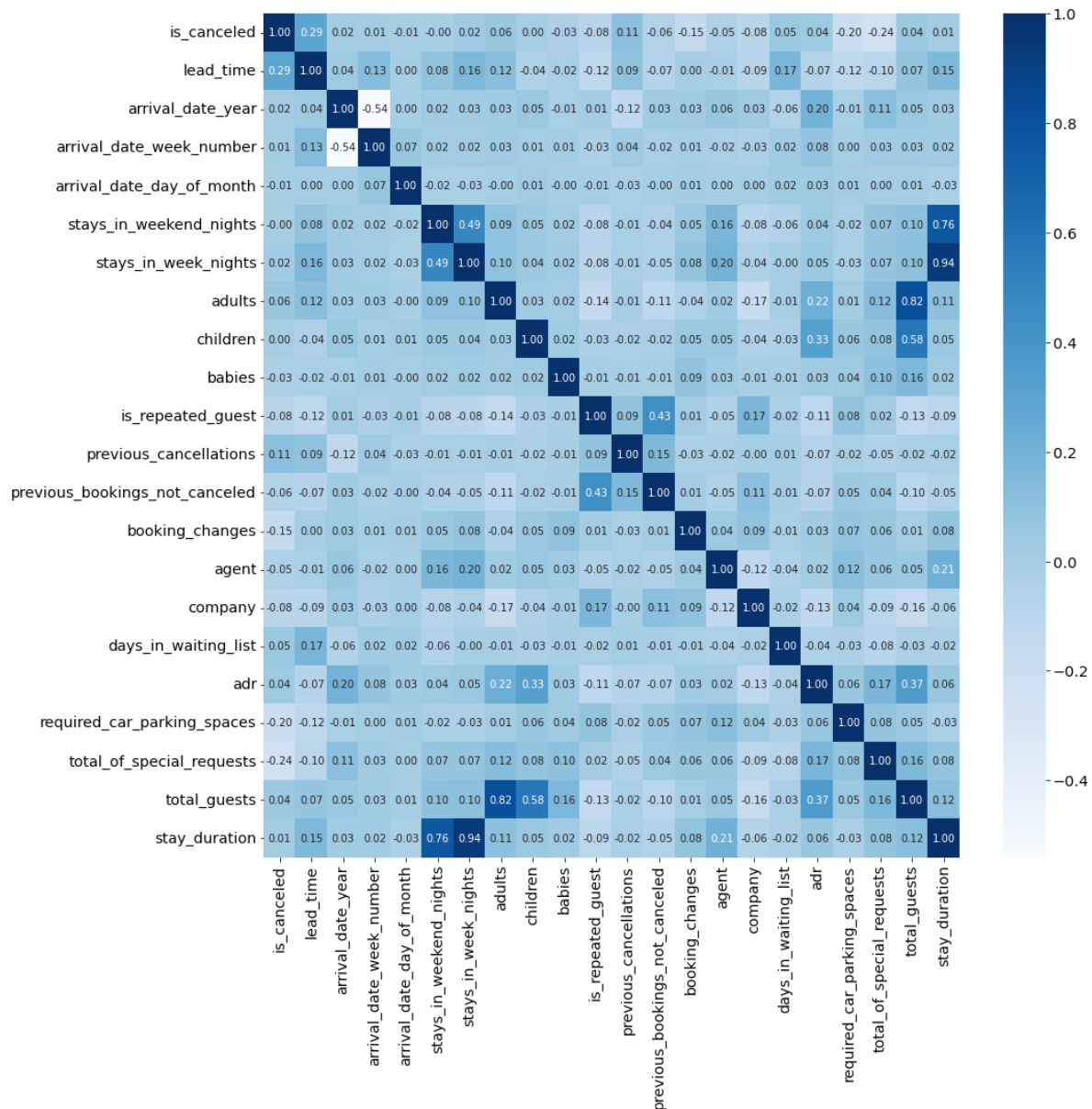


Dari 3 hasil plot di atas, dapat dilihat bahwa

- Kolom “stays\_in\_week\_nights”, “stays\_in\_weekend\_nights”, “stay\_duration” mempunyai distribusi positively skewed. (Melihat bentuk box plot, distribution plot, dan violin plot). Langkah selanjutnya adalah melakukan menormalkan distribusi data, metode yang digunakan bisa menggunakan Log transformation, dll. Setelah distribusi data menjadi normal, selanjutnya kita lakukan feature scaling bisa berupa Min-max scaler, standard scaler, dan robust scaler, dll.
- Terdapat beberapa kolom yang memiliki outliers, seperti kolom “total\_of\_special\_requests”, “required\_car\_parking\_spaces”, “children”, “babies”, “adults”. Dari box plot, terdapat beberapa jenis outlier seperti global outlier dan collective outlier. Untuk pemodelan, biasanya kita membuang global outlier (outlier yang sangat ekstrem), sementara kita membiarkan collective outlier tetap ada.
- Terdapat beberapa nilai yang mendominasi, contohnya nilai 0 pada kolom “babies”, 0 pada kolom children, “stay duration” yang mendominasi tidak lebih dari 10 hari, dll.

## 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?

Jawab:

Untuk korelasi antara feature dan label (“is\_cancelled”) tidak memiliki korelasi yang kuat. Adapun feature yang bertipe numerik ternyata berupa ID, yaitu “Agent”, “company”. Kolom tersebut direkomendasikan untuk didrop sementara feature yang lain masih relevan untuk eksplorasi lebih lanjut.

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

**Jawab:**

Dari correlation heatmap, kita dapat memperoleh korelasi antar feature. Jika korelasi  $> 0.7$  artinya terdapat hubungan 2 feature yang menandakan multikolinieritas/redundan sehingga salah satu feature harus dibuang. Pada diagram heatmap diperoleh bahwa fitur “stay duration” berkorelasi  $> 0.7$  terhadap fitur “stay in weekend nights” dan “stay in week nights”. Selain itu, feature “total guest” juga berkorelasi  $> 0.7$  dengan feature “adults” dan “children”.

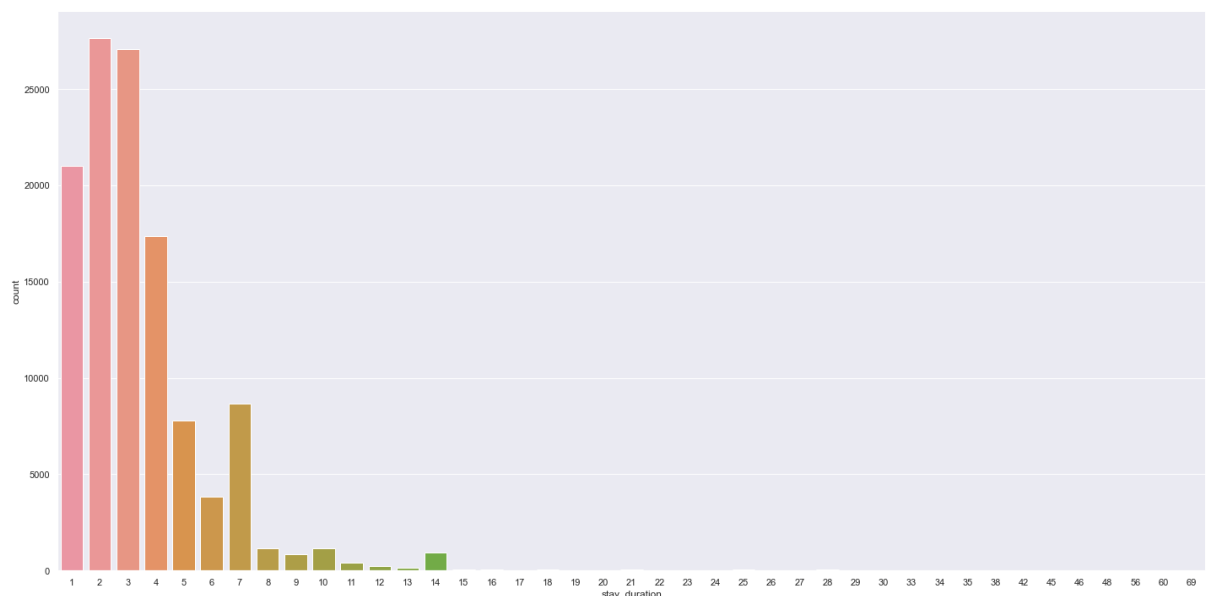
#### 4. Business Insight

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.

**Jawab:**

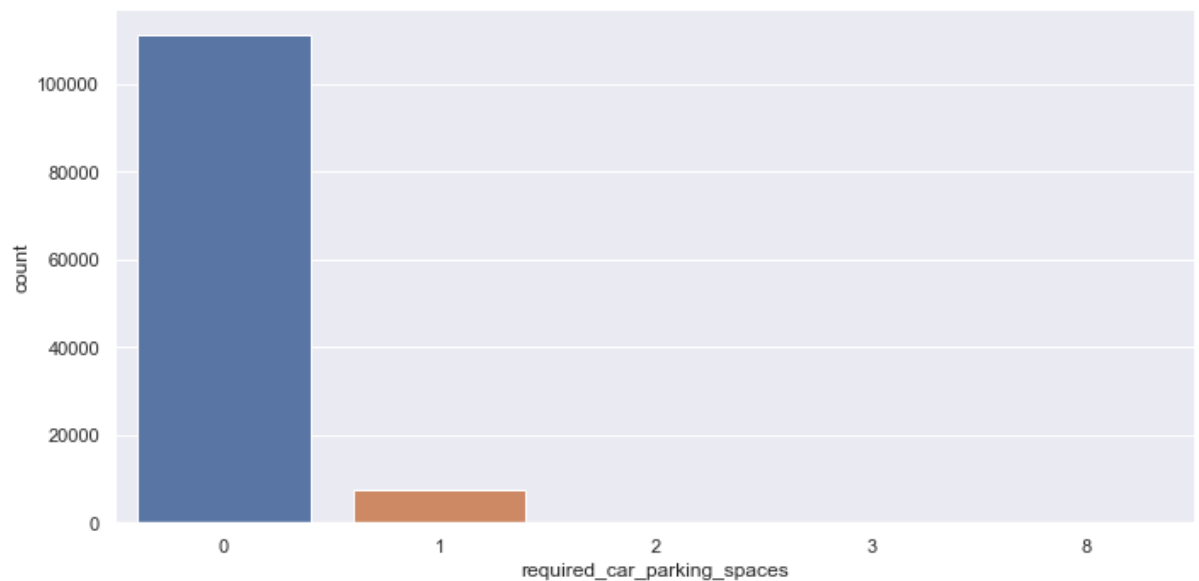
Ada beberapa insight yang dapat diambil seperti:

- a. Pelanggan kebanyakan stay di hotel dengan durasi stay kurang dari 10 hari



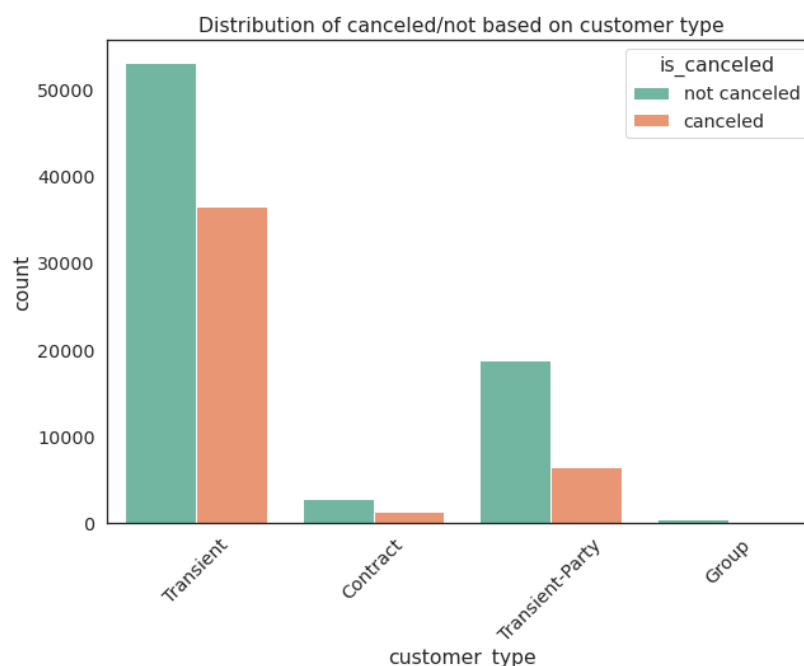
Dari hal ini pihak hotel dapat memfokuskan usaha pada pelanggan yang melakukan short stay duration saja. Misal, melakukan promo tambahan breakfast dan lunch pada pelanggan yang hanya stay selama 2 hari dengan tambahan biaya, sehingga dapat memaksimalkan profit.

b. Pelanggan tidak banyak menggunakan mobil untuk stay di hotel



Hal ini dapat menjadi peluang bisnis bagi pihak hotel untuk menyediakan moda transportasi untuk pelanggan hotel yang ingin berwisata, maupun pergi menuju ke airport, atau bandara (Bisa menyediakan jasa sewa sopir)

c. Distribusi banyaknya pelanggan yang canceled dan yang tidak berdasarkan tipe pelanggan.



Dari visualisasi tersebut diperoleh bahwa customer bertipe "Transient" (sementara) merupakan paling banyak membooking hotel, disisi lain tipe customer tersebut juga paling banyak melakukan cancel booking hotel. Pihak hotel lebih baik fokus kepada tipe customer lain, karena customer tipe transient belum pasti melakukan booking di hotel.



## 5. Git

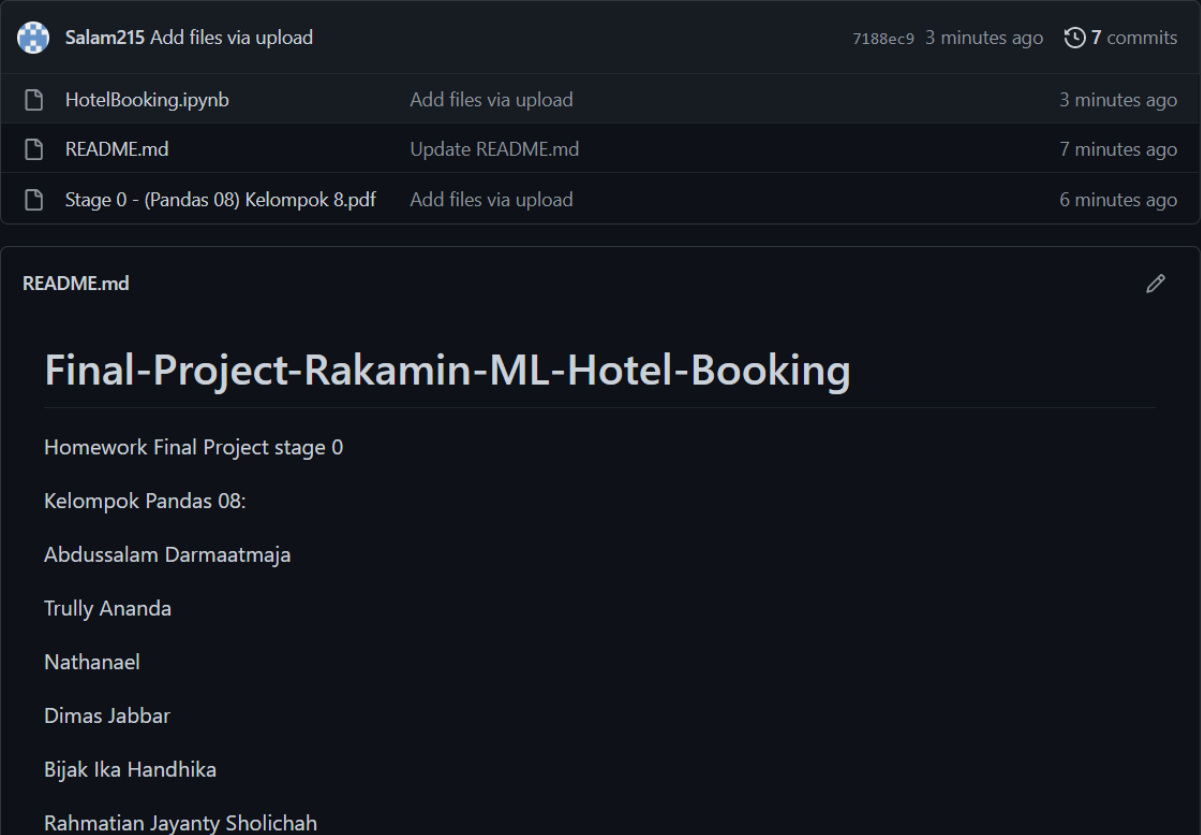
Upload project teman-teman di sebuah repository git. Berkolaborasi Lah diGit jika ada perubahan version dari waktu ke waktu.

### A. Buat Repository Git

Link GitHub untuk Final Projek

<https://github.com/Salam215/Final-Project-Rakamin-ML-Hotel-Booking>

B. Upload file notebook atau file pengerjaan lainnya pada repository tersebut Untuk file README, dapat merupakan summary insight yang telah didapatkan dari EDA.



The screenshot shows a GitHub repository interface. At the top, the repository name 'Salam215' is displayed with a globe icon, followed by the text 'Add files via upload'. To the right, the commit hash '7188ec9' is shown, along with '3 minutes ago' and '7 commits'. Below this, a table lists recent file uploads:

File Name	Action	Time
HotelBooking.ipynb	Add files via upload	3 minutes ago
README.md	Update README.md	7 minutes ago
Stage 0 - (Pandas 08) Kelompok 8.pdf	Add files via upload	6 minutes ago

Below the table, the 'README.md' file is selected, showing its content. The title is 'Final-Project-Rakamin-ML-Hotel-Booking'. The content includes:

- Homework Final Project stage 0
- Kelompok Pandas 08:
- Abdussalam Darmaatmaja
- Trully Ananda
- Nathanael
- Dimas Jabbar
- Bijak Ika Handhika
- Rahmatian Jayanty Sholichah