Rakamin Academy

Fantastic Four

Laporan Final Project Stage - 1

(dipresentasikan setiap sesi mentoring)

Rakamin Academy

Anggota Kelompok :

Farha Apita

Hafizh M R

Quini Arantxa

Rafly Syafiq

Refia K

Siswah

Table Of Contents :

Project Background

Dataset Hotel_Booking	
Problem	•Sekitar 37% hotel booking ternyata cancelled, perlu diminimalisasi
Role	•Data Scientist of Hotel industry
Goal	•Menurunkan jumlah pembatalan reservasi kamar hotel
Objective	Memprediksi pelanggan yang berisiko membatalkan hotel booking dan membuat kebijakan baru agar pelanggan mengurungkan pembatalan sehingga diharapkan menurunkan cancellation rate
Business metrics	Cancellation rate (seberapa besar customer yang melakukan cancel booking)

EDA (Exploratory Data Analysis)

Descriptive Statistics

Rakamin Academy

Numerical Describe

```
[ ] central = ['mean', '25%', '50%', '75%', 'mode']
    spread = ['min', 'max', 'range', 'std', 'variance', 'IQR']

    df_nums[['count', 'unique'] + central + spread]
```

	count	unique	mean	25%	50%	75%	mode	min	max	range	std	variance	IQR
is_canceled	119390.0	2	0.370416	0.00	0.000	1.0	0.0	0.00	1.0	1.00	0.482918	0.23	1.00
lead_time	119390.0	479	104.011416	18.00	69.000	160.0	0.0	0.00	737.0	737.00	106.863097	11419.72	142.00
arrival_date_year	119390.0	3	2016.156554	2016.00	2016.000	2017.0	2016.0	2015.00	2017.0	2.00	0.707476	0.50	1.00
arrival_date_week_number	119390.0	53	27.165173	16.00	28.000	38.0	33.0	1.00	53.0	52.00	13.605138	185.10	22.00
arrival_date_day_of_month	119390.0	31	15.798241	8.00	16.000	23.0	17.0	1.00	31.0	30.00	8.780829	77.10	15.00
stays_in_weekend_nights	119390.0	17	0.927599	0.00	1.000	2.0	0.0	0.00	19.0	19.00	0.998613	1.00	2.00
stays_in_week_nights	119390.0	35	2.500302	1.00	2.000	3.0	2.0	0.00	50.0	50.00	1.908286	3.64	2.00
adults	119390.0	14	1.856403	2.00	2.000	2.0	2.0	0.00	55.0	55.00	0.579261	0.34	0.00
children	119386.0	5	0.103890	0.00	0.000	0.0	0.0	0.00	10.0	10.00	0.398561	0.16	0.00
babies	119390.0	5	0.007949	0.00	0.000	0.0	0.0	0.00	10.0	10.00	0.097436	0.01	0.00
is_repeated_guest	119390.0	2	0.031912	0.00	0.000	0.0	0.0	0.00	1.0	1.00	0.175767	0.03	0.00
previous_cancellations	119390.0	15	0.087118	0.00	0.000	0.0	0.0	0.00	26.0	26.00	0.844336	0.71	0.00
previous_bookings_not_canceled	119390.0	73	0.137097	0.00	0.000	0.0	0.0	0.00	72.0	72.00	1.497437	2.24	0.00
booking_changes	119390.0	21	0.221124	0.00	0.000	0.0	0.0	0.00	21.0	21.00	0.652306	0.43	0.00
agent	103050.0	333	86.693382	9.00	14.000	229.0	9.0	1.00	535.0	534.00	110.774548	12271.00	220.00
company	6797.0	352	189.266735	62.00	179.000	270.0	40.0	6.00	543.0	537.00	131.655015	17333.04	208.00
days_in_waiting_list	119390.0	128	2.321149	0.00	0.000	0.0	0.0	0.00	391.0	391.00	17.594721	309.57	0.00
adr	119390.0	8879	101.831122	69.29	94.575	126.0	62.0	-6.38	5400.0	5406.38	50.535790	2553.87	56.71
required_car_parking_spaces	119390.0	5	0.062518	0.00	0.000	0.0	0.0	0.00	8.0	8.00	0.245291	0.06	0.00
total_of_special_requests	119390.0	6	0.571363	0.00	0.000	1.0	0.0	0.00	5.0	5.00	0.792798	0.63	1.00

Descriptive Statistics

Rakamin Academy

Categorical Describe

```
[ ] df_cats = df[cats].describe().T
    df_cats
```

	count	unique	top	freq
hotel	119390	2	City Hotel	79330
arrival_date_month	119390	12	August	13877
meal	119390	5	BB	92310
country	118902	177	PRT	48590
market_segment	119390	8	Online TA	56477
distribution_channel	119390	5	TA/TO	97870
reserved_room_type	119390	10	A	85994
assigned_room_type	119390	12	A	74053
deposit_type	119390	3	No Deposit	104641
customer_type	119390	4	Transient	89613
reservation_status	119390	3	Check-Out	75166
reservation_status_date	119390	926	2015-10-21	1461
name	119390	81503	Michael Johnson	48
email	119390	115889	Michael.C@gmail.com	6
phone-number	119390	119390	669-792-1661	1
credit_card	119390	9000	**********4923	28
countries	119390	176	Portugal	48590

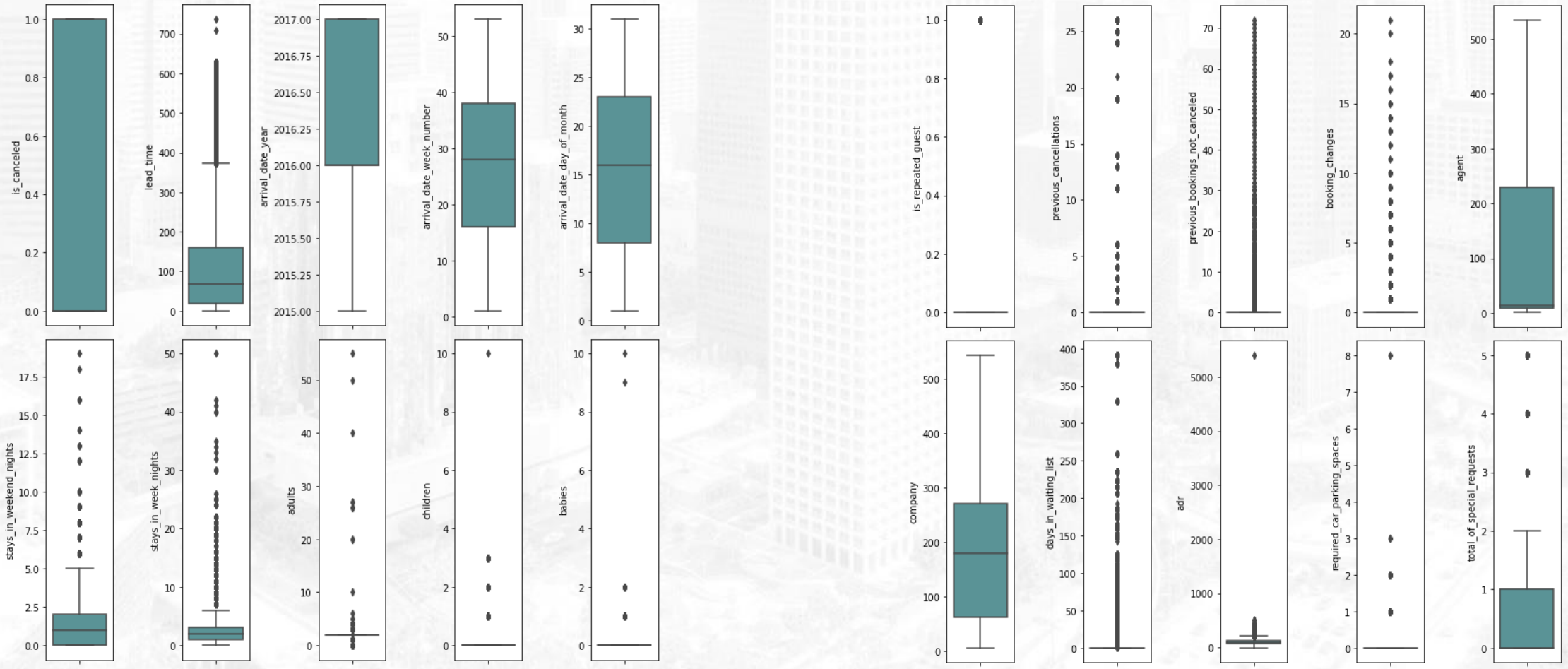
Insight Descriptive Statistics

1. Tampak beberapa kolom, yakni kolom childern, country, agent, dan company masih memiliki null/missing values (Non-Null Count < jumlah baris)
2. Pada kolom company ditemukan jumlah null/missing values yang sangat banyak, dimana dari 119390 baris ditemukan jumlah baris yang non-null hanya 6797
3. Pada kolom children, tipe data float64 kurang sesuai karena kolom tersebut menunjukkan jumlah anak (number of childern). Tipe data yang sesuai seharusnya int64
4. Pada kolom company dan agent, tipe data float64 juga kurang sesuai karena kolom tersebut menunjukkan ID. Tipe data yang lebih baik adalah object
5. Pada kolom reservation_status_date bertipe object, seharusnya bertipe datetime

Univariate Analysis

Boxplots (Numeric)

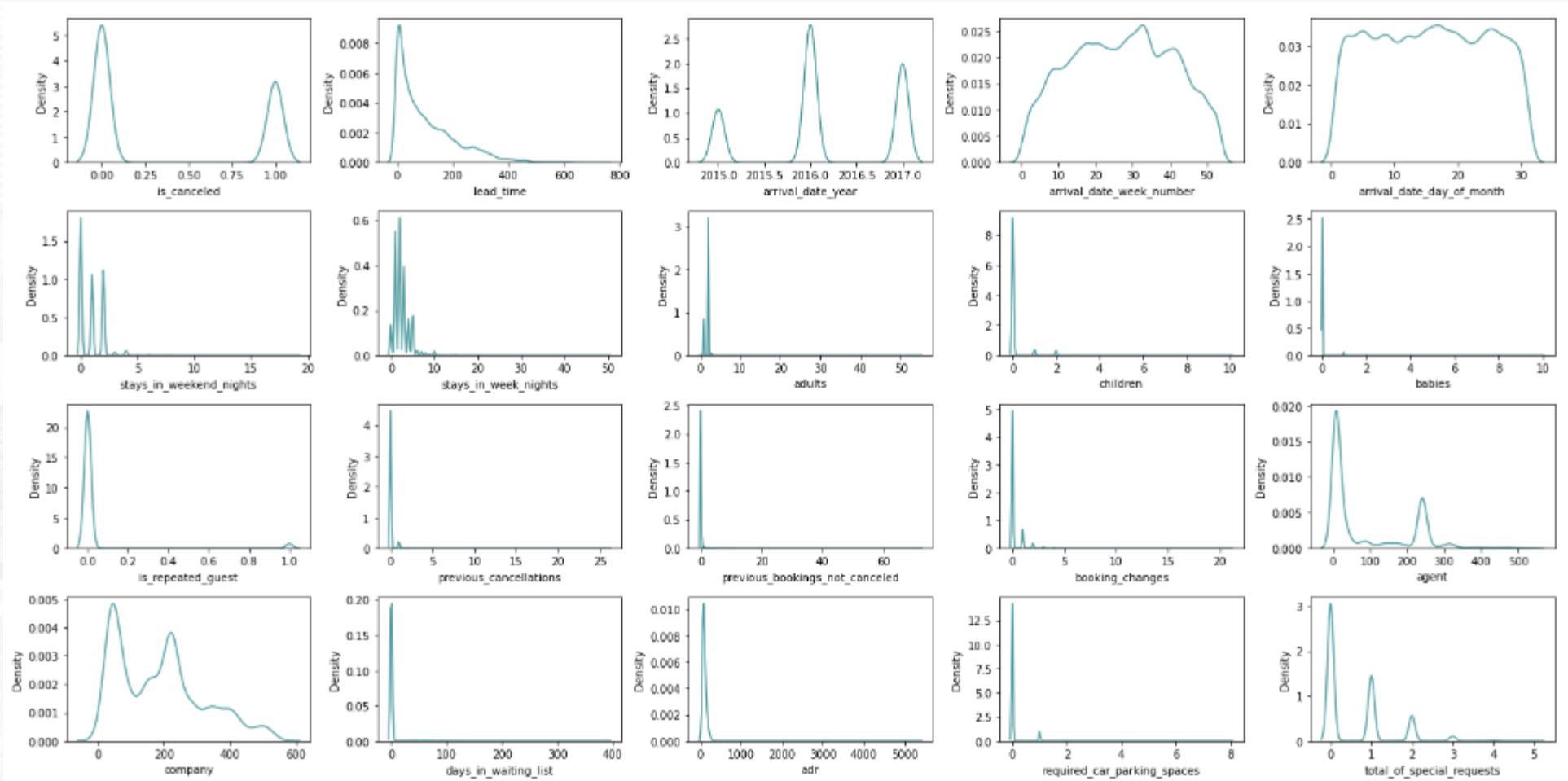
```
features = numericals
plt.figure(figsize=(10, 20))
for i in range(0, len(features)):
    plt.subplot(4, 5, i+1)
    sns.boxplot(y=df[features[i]], color='#509ca0', orient='v')
    plt.tight_layout();
```



Univariate Analysis

- Distplot (Numerical)

```
[ ]  features = numericals
     plt.figure(figsize=(20, 10))
     for i in range(0, len(features)):
         plt.subplot(4, 5, i+1)
         sns.kdeplot(x=df[features[i]], color='#509ca0')
         plt.xlabel(features[i])
         plt.tight_layout();
```

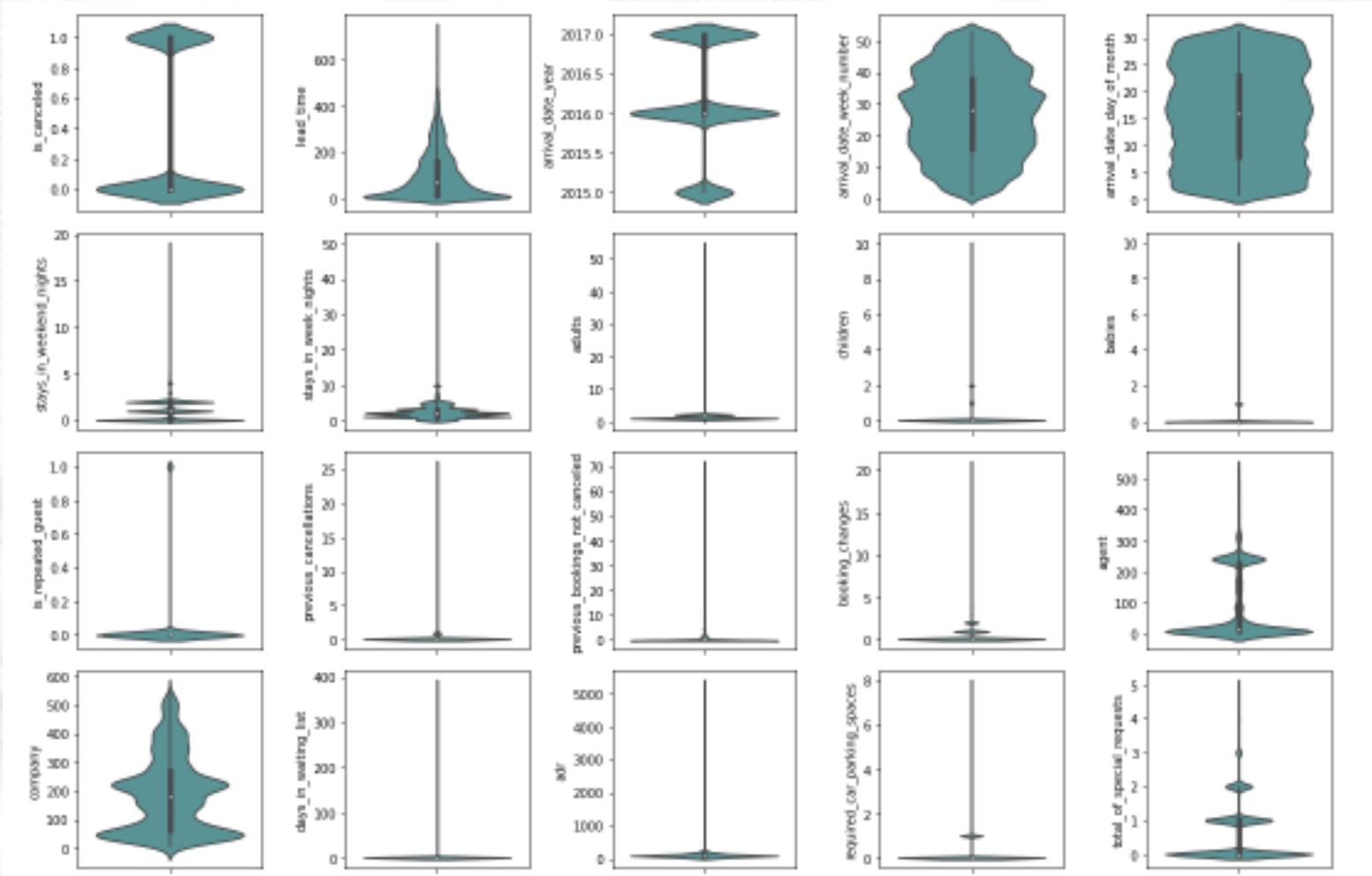


Univariate Analysis

Rakamin Academy

▼ Violin Plots (Numeric)

```
features = numericals
plt.figure(figsize=(15, 15))
for i in range(0, len(features)):
    plt.subplot(4, 5, i+1)
    sns.violinplot(y=df[features[i]], color='#509ca0')
    plt.tight_layout();
```

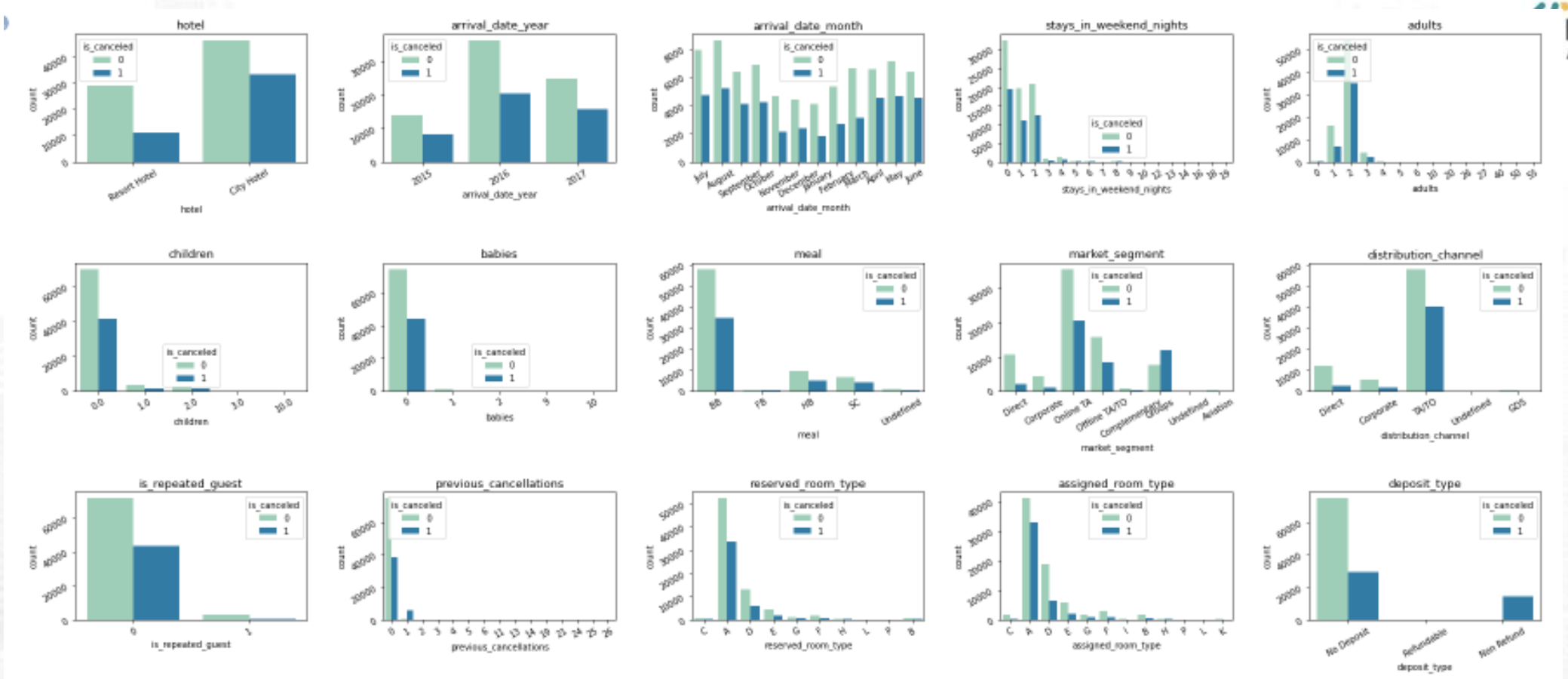


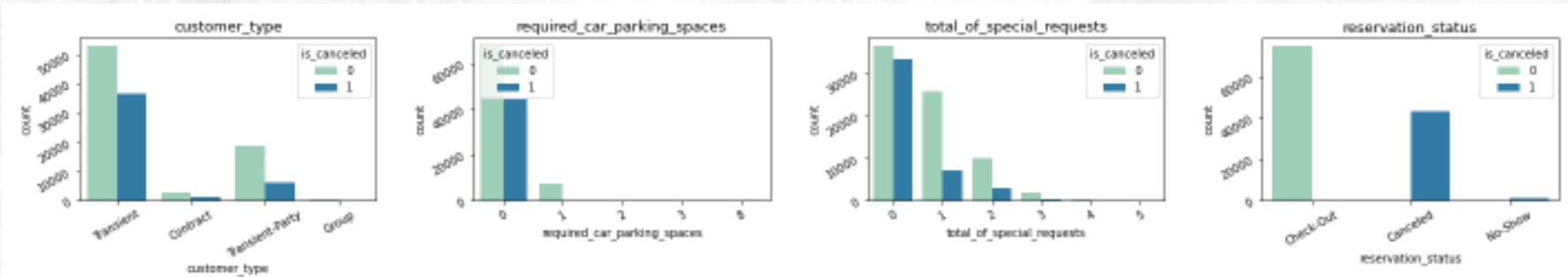
Univariate Analysis

Rakamin Academy

- Countplot (Categorical)

```
[ ]  d_visual = df.nunique()[df.nunique()<20].drop(["is_canceled"]).index
     fig = plt.figure(figsize=(25,15))
     for index, col in enumerate(d_visual):
         ax = fig.add_subplot(4, 5, index+1)
         ax.set_title(col,fontsize=13)
         ax.tick_params(labelrotation=30)
         sns.countplot(df[col], hue=df.is_canceled, ax=ax, palette="YlGnBu")
     plt.tight_layout(pad=3);
```

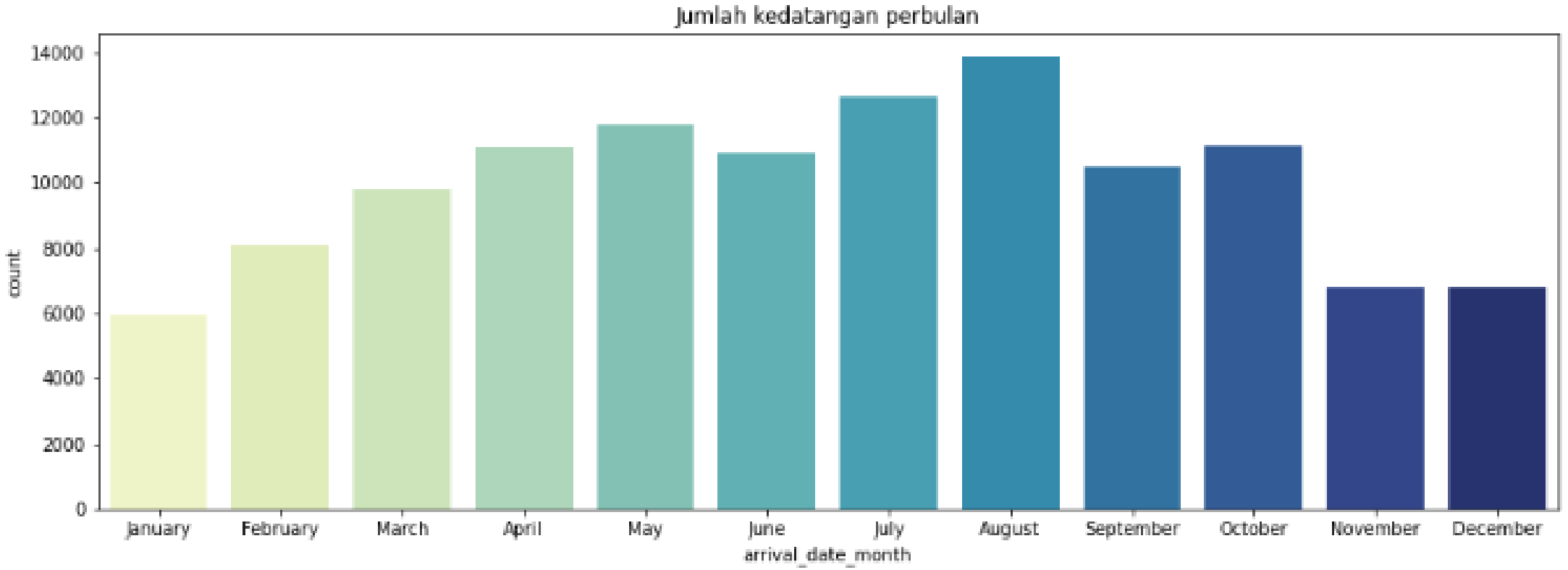

hotel
arrival_date_year
arrival_date_month
stays_in_weekend_nights
adults
children
babies
meal
market_segment
distribution_channel
is_repeated_guest
previous_cancellations
reserved_room_type
assigned_room_type
deposit_type
is_canceled
count


customer_type
required_car_parking_spaces
total_of_special_requests
reservation_status
is_canceled
count

Univariate Analysis

Rakamin Academy

```
[ ]  plt.figure(figsize=(15, 5));
     plt.title("Jumlah kedatangan perbulan");
     sns.countplot(df.arrival_date_month, palette="YlGnBu", order=calendar.month_name[1:]);
```

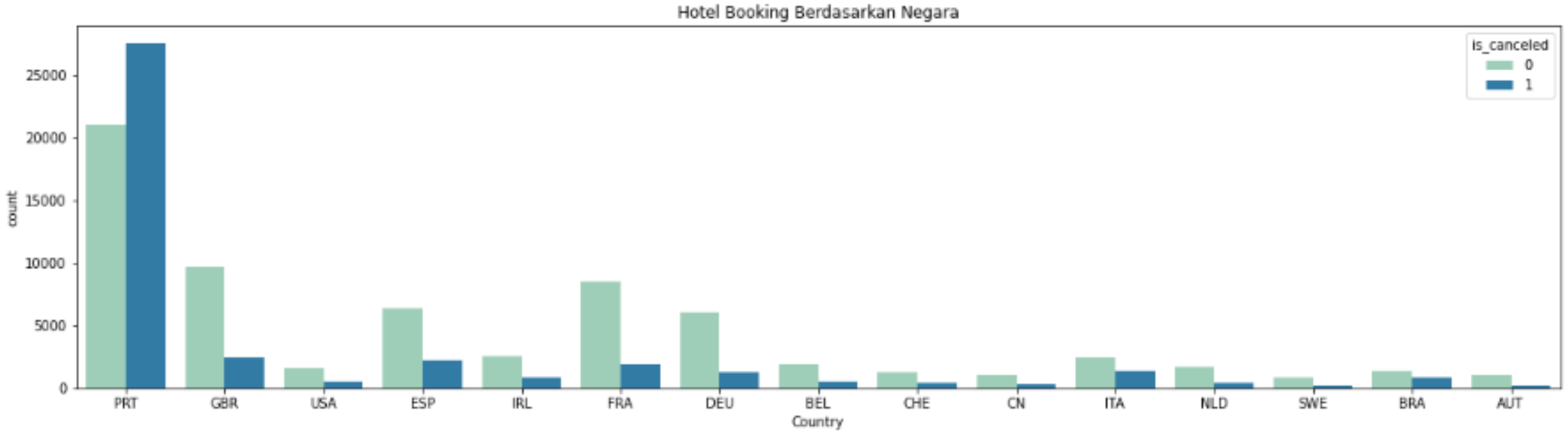


Univariate Analysis

Rakamin Academy

```
[ ]  df_country = df['country'].value_counts().sort_values(ascending=False)[:15]
     plt.figure(figsize=(20,5))
     sns.countplot(x='country', hue='is_canceled', data=df[df['country'].isin(df_country.index)], palette="YlGnBu")
     plt.xlabel("Country")
     plt.title("Hotel Booking Berdasarkan Negara")
```

```
Text(0.5, 1.0, 'Hotel Booking Berdasarkan Negara')
```



Insight Univariate Analysis

1. Berdasarkan visualisasi countplot, pada kolom hotel terlihat bahwa City Hotel memiliki jumlah cancellation tertinggi.
2. Pada bagian boxplot, beberapa variable memiliki data yang cenderung miring ke kanan (skewness positif) dan beberapa cenderung miring ke kiri (skewness kiri).
3. Beberapa variabel memiliki data outlier yang harus ditindaklanjut, seperti kolom *lead_time, stays_in_weekend_nights, stays_in_week_nights, adults, children, babies, previous_cancellation, previous_boking_not_canceled, booking_changes, days_in_waiting_list, adr, required_car_parking_lot, total_of_special_request*
4. Bila dilihat secara visual pada Distplot, variabel arrival_date_week_number memiliki distribusi normal

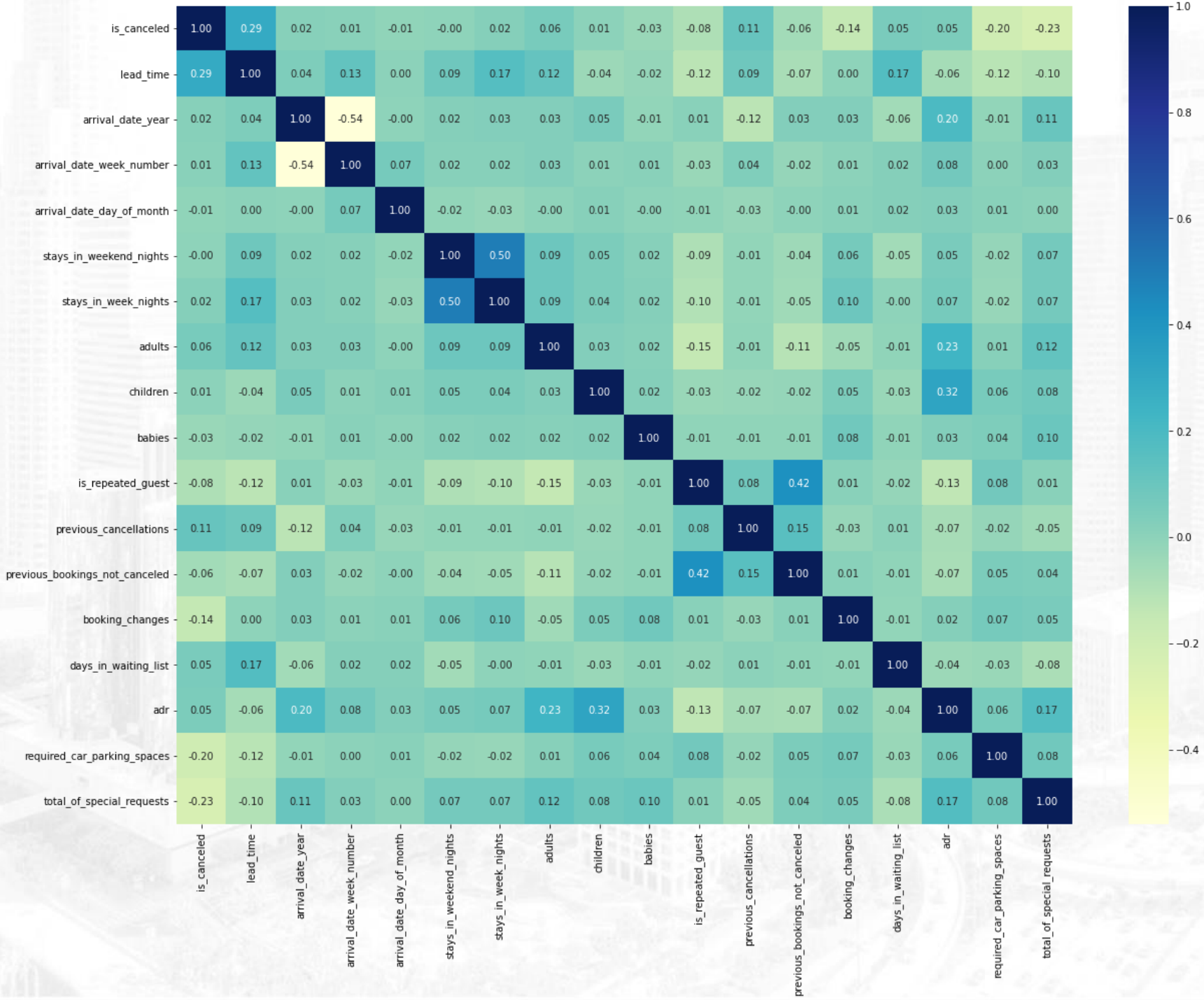
Multivariate Analysis

Rakamin Academy

- Multivariate Analysis

- Correlation Heatmap

```
df2 = df.drop(['agent', 'company'], axis=1)
numericals2 = df2.loc[:, (df2.dtypes == int) | (df2.dtypes == float)].columns.tolist()
plt.figure(figsize=(20,15))
sns.heatmap(df2[numericals2].corr(), cmap='YlGnBu', annot=True, fmt='.2f');
```

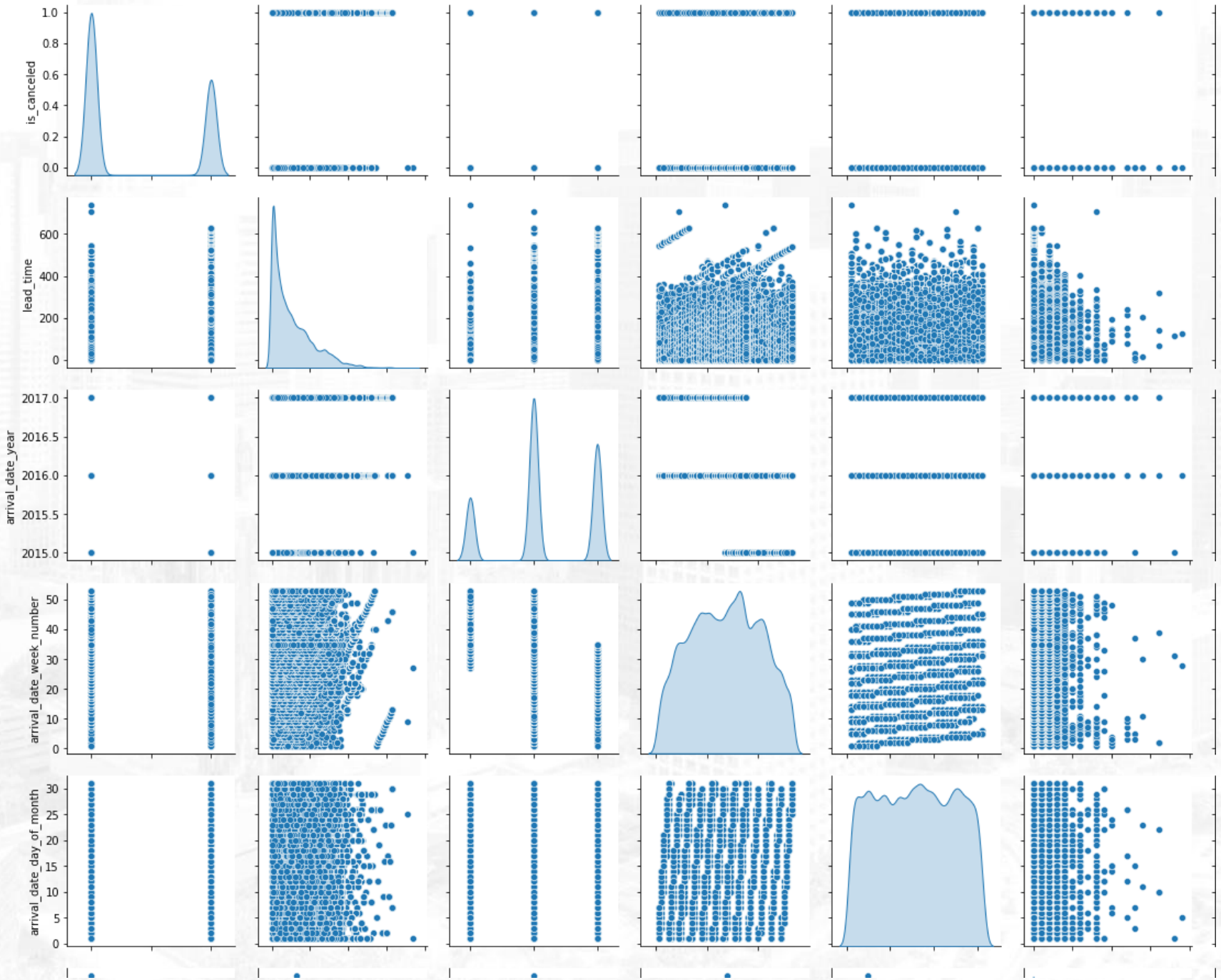

is_canceled 1.00 0.29 0.02 0.01 -0.01 -0.00 0.02 0.06 0.01 -0.03 -0.08 0.11 -0.06 -0.14 0.05 0.05 -0.20 -0.23
lead_time 0.29 1.00 0.04 0.13 0.00 0.09 0.17 0.12 -0.04 -0.02 -0.12 0.09 -0.07 0.00 0.17 -0.06 -0.12 -0.10
arrival_date_year 0.02 0.04 1.00 -0.54 -0.00 0.02 0.03 0.03 0.05 -0.01 0.01 -0.12 0.03 0.03 -0.06 0.20 -0.01 0.11
arrival_date_week_number 0.01 0.13 -0.54 1.00 0.07 0.02 0.02 0.03 0.01 0.01 -0.03 0.04 -0.02 0.01 0.02 0.08 0.00 0.03
arrival_date_day_of_month -0.01 0.00 -0.00 0.07 1.00 -0.02 -0.03 -0.00 0.01 -0.00 -0.01 -0.03 -0.00 0.01 0.02 0.03 0.01 0.00
stays_in_weekend_nights -0.00 0.09 0.02 0.02 -0.02 1.00 0.50 0.09 0.05 0.02 -0.09 -0.01 -0.04 0.06 -0.05 0.05 -0.02 0.07
stays_in_week_nights 0.02 0.17 0.03 0.02 -0.03 0.50 1.00 0.09 0.04 0.02 -0.10 -0.01 -0.05 0.10 -0.00 0.07 -0.02 0.07
adults 0.06 0.12 0.03 0.03 -0.00 0.09 0.09 1.00 0.03 0.02 -0.15 -0.01 -0.11 -0.05 -0.01 0.23 0.01 0.12
children 0.01 -0.04 0.05 0.01 0.01 0.05 0.04 0.03 1.00 0.02 -0.03 -0.02 -0.02 0.05 -0.03 0.32 0.06 0.08
babies -0.03 -0.02 -0.01 0.01 -0.00 0.02 0.02 0.02 0.02 1.00 -0.01 -0.01 -0.01 0.08 -0.01 0.03 0.04 0.10
is_repeated_guest -0.08 -0.12 0.01 -0.03 -0.01 -0.09 -0.10 -0.15 -0.03 -0.01 1.00 0.08 0.42 0.01 -0.02 -0.13 0.08 0.01
previous_cancellations 0.11 0.09 -0.12 0.04 -0.03 -0.01 -0.01 -0.01 -0.02 -0.01 0.08 1.00 0.15 -0.03 0.01 -0.07 -0.02 -0.05
previous_bookings_not_canceled -0.06 -0.07 0.03 -0.02 -0.00 -0.04 -0.05 -0.11 -0.02 -0.01 0.42 0.15 1.00 0.01 -0.01 -0.07 0.05 0.04
booking_changes -0.14 0.00 0.03 0.01 0.01 0.06 0.10 -0.05 0.05 0.08 0.01 -0.03 0.01 1.00 -0.01 0.02 0.07 0.05
days_in_waiting_list 0.05 0.17 -0.06 0.02 0.02 -0.05 -0.00 -0.01 -0.03 -0.01 -0.02 0.01 -0.01 -0.01 1.00 -0.04 -0.03 -0.08
adr 0.05 -0.06 0.20 0.08 0.03 0.05 0.07 0.23 0.32 0.03 -0.13 -0.07 -0.07 0.02 -0.04 1.00 0.06 0.17
required_car_parking_spaces -0.20 -0.12 -0.01 0.00 0.01 -0.02 -0.02 0.01 0.06 0.04 0.08 -0.02 0.05 0.07 -0.03 0.06 1.00 0.08
total_of_special_requests -0.23 -0.10 0.11 0.03 0.00 0.07 0.07 0.12 0.08 0.10 0.01 -0.05 0.04 0.05 -0.08 0.17 0.08 1.00
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
children
babies
is_repeated_guest
previous_cancellations
previous_bookings_not_canceled
booking_changes
days_in_waiting_list
adr
required_car_parking_spaces
total_of_special_requests
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Multivariate Analysis

Rakamin Academy

Pair Plots (Numeric)

```
plt.figure(figsize=(15, 15))
sns.pairplot(df[numericals], diag_kind='kde', palette='#509ca0')
plt.show()
```


is_canceled
lead_time
arrival_date_year
arrival_date_week_number
arrival_date_day_of_month

Insight Multivariate Analysis

1. Terlihat Target memiliki korelasi positif dengan fitur adr, days_in_waiting_list, previous_cancellations, children, adults, stays_in_weeks_nights, arrival_date_week_number, arrival_date_year, lead_time
2. Tidak ada variabel yang berpotensi menyebabkan multicollinearity.
3. stays_in_weeks_night dan stays_in_weekend_night memiliki nilai korelasi paling tinggi yaitu 0,5.

Thank You ☺!