



Hotel Booking

Nguyen Thi Lan Thanh

PRESENTATION OUTLINE

Overview

01

EDA

02

Machine Learning

03

Conclusion

04

Recommendation

05

Introduction

- This dataset contains 118,898 observations for a City Hotel and a Resort Hotel. Each observation represents a hotel booking between the July 2015 and July 2017, including booking that effectively arrived and booking that were canceled.
- Dataset Hotel booking at [LINK](#)
- Data: 118,898 rows x 17 columns

Context

- **You are a data analyst** at a resort/hotel. You have this dataset and your task is to analyze it in order to help reduce the booking cancellation rate for the resort/hotel.
- **Report to:** Marketing, Sales

Goal

- **Reduce the booking cancellation rate** for the resort/hotel.

[Overview](#)[EDA](#)[Machine Learning](#)[Conclution](#)[Recommendation](#)

Data Tools



PYTHON

Mining, Cleaning,
Observing and
Calculating Dataset.



POWERBI

Visualize the data
and building
Dashboard

Overview

EDA

Machine Learning

Conclution

Recommendation

Column Name	Discription
hotel	hotel: Type of accommodation (e.g., hotel or resort).
is_canceled	is_canceled: Booking cancellation status (1: Canceled, 0: Not canceled).
lead_time	lead_time: Number of days between the booking date and the check-in date.
arrival_date_year	arrival_date_year: Year of guest arrival.
arrival_date_month	arrival_date_month: Month of guest arrival.
arrival_date_day_of_month	arrival_date_day_of_month: Day of the month the guest arrives.
stays_in_weekend_nights	stays_in_weekend_nights: Number of weekend nights stayed (Saturday and Sunday).
stays_in_week_nights	stays_in_week_nights: Number of weekday nights stayed (Monday to Friday).
adults	adults: Number of adults included in the booking.
children	children: Number of children included in the booking.
babies	babies: Number of babies included in the booking.
country	country: Country of origin of the guest.
market_segment	market_segment: Market segment (e.g., Online Travel Agency, Offline Travel Agency, Direct, Corporate, etc.).
is_repeated_guest	is_repeated_guest: Whether the guest is a returning customer (1: Yes, 0: No).
customer_type	customer_type: Type of customer (e.g., Transient, Contract, Group, Transient-Party).
adr	adr: Average Daily Rate (ADR) of the booking.
total_of_special_requests	total_of_special_requests: Total number of special requests made by the guest (e.g., extra bed, non-smoking room, etc.).

HOTEL CANCELLATION RATE ANALYSIS

Total
118.90K

Cancellation
44.15K

Cancellation Rate
37.14%

Repeated Rate
3.20%

Year

☒ 2015

☒ 2016

☒ 2017

Hotel

City Hotel

Resort Hotel

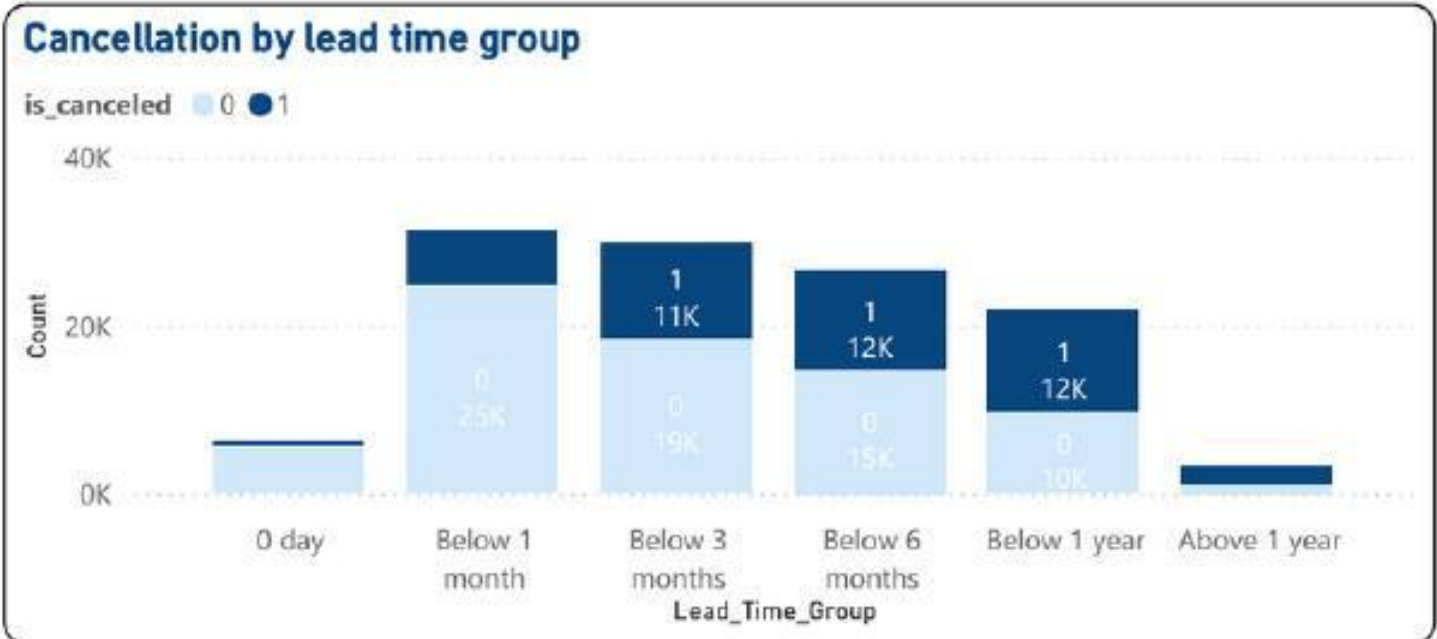
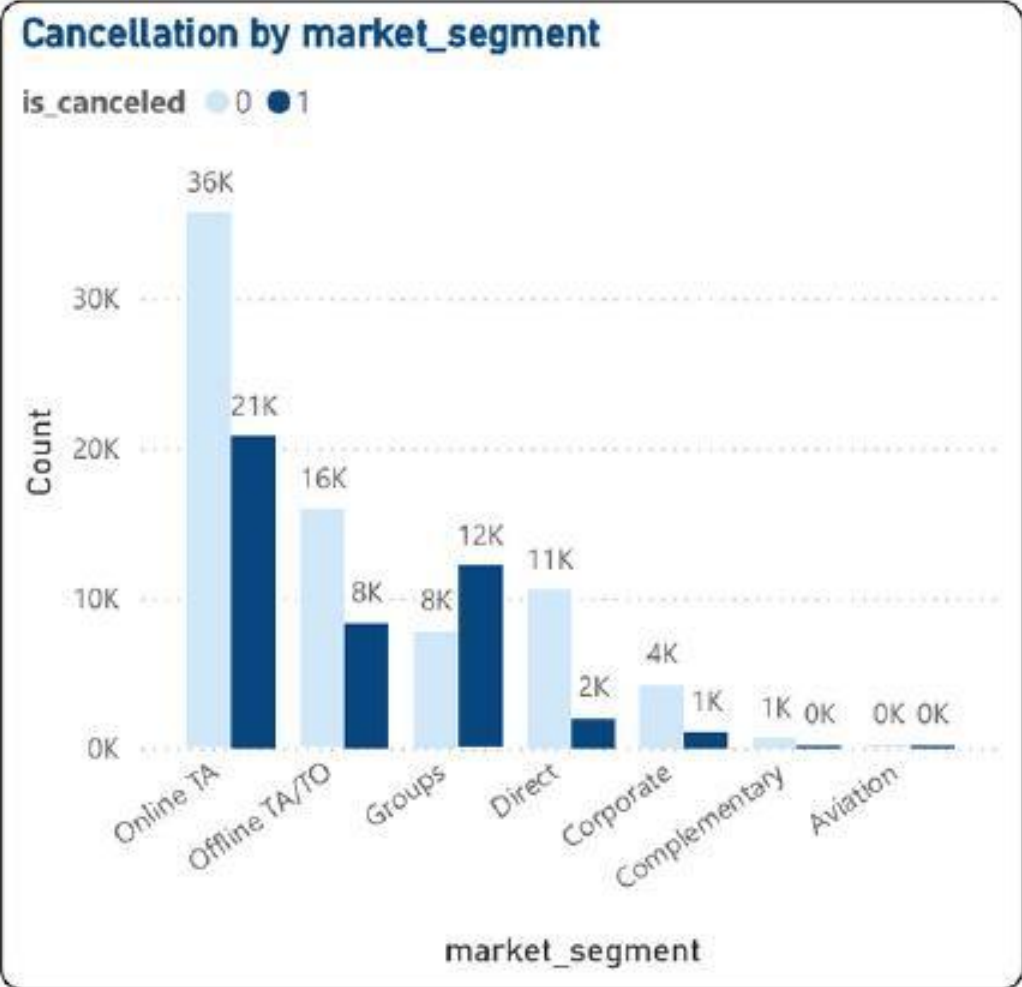
Customer type

Contract

Group

Transient

Transient-Party





118.90K

Total Booking

44.15K

Cancellation

37.14%

Cancellation Rate

The cancellation rate is 37.14% → a relatively high proportion
→ Analyze the characteristics of canceled bookings to develop appropriate business strategies

Overview

EDA

Machine Learning

Conclution

Recommendation

City Hotel

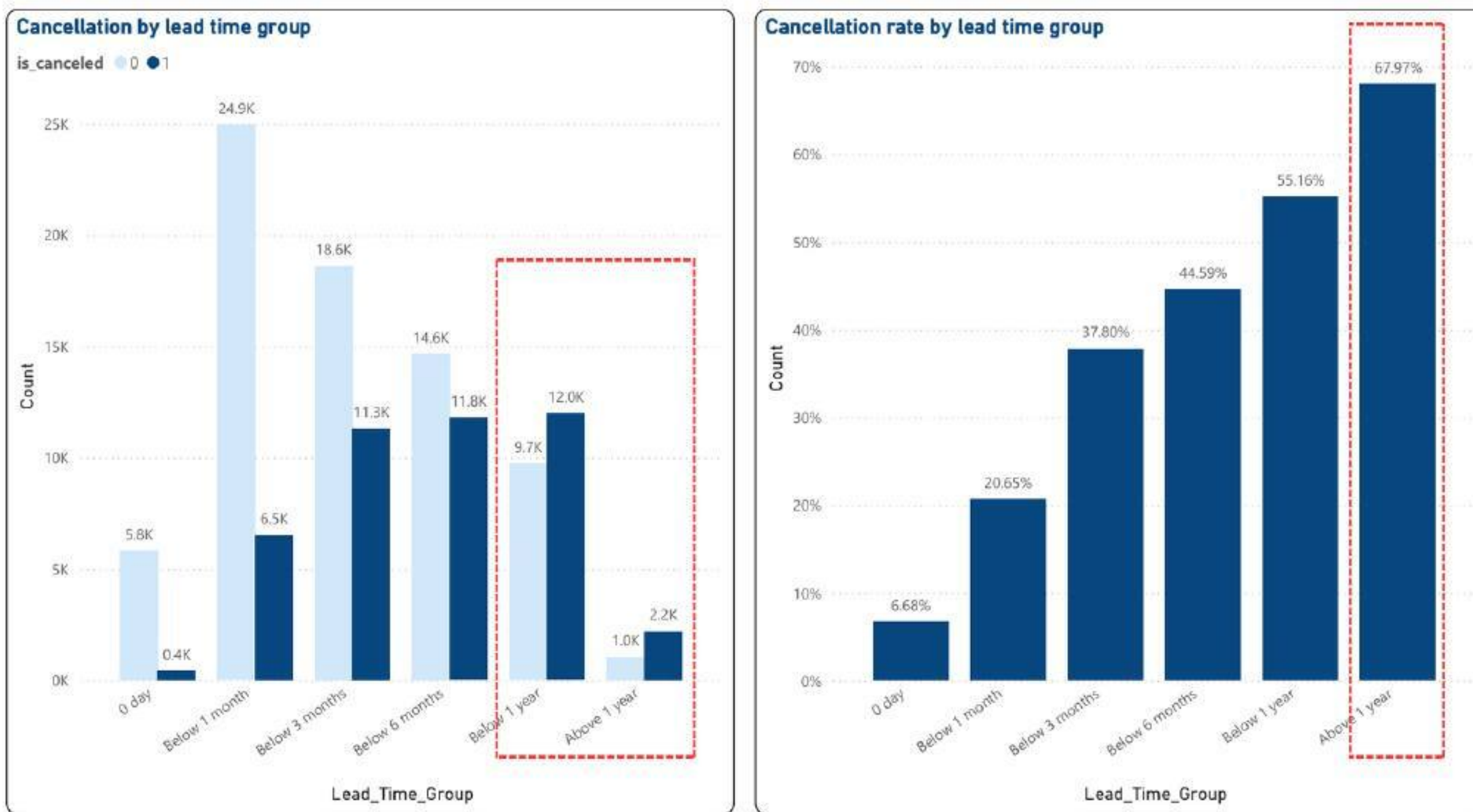


- The cancellation rate is **41.71% higher than the average**.
- As lead time increases, the cancellation rate also rises.
- The cancellation rate consistently stays above 35% across all time periods, peaking in July 2015 at 50.11%, followed by **continuous fluctuations**.
- The highest cancellation rate is in the **“Groups” segment (68.86%)**; The **“Transient” and “Contract” customer types almost 50%**
- Customers **without special requests** have a very **high cancellation rate** (nearly 55%), which gradually decreases as the number of special requests increases.

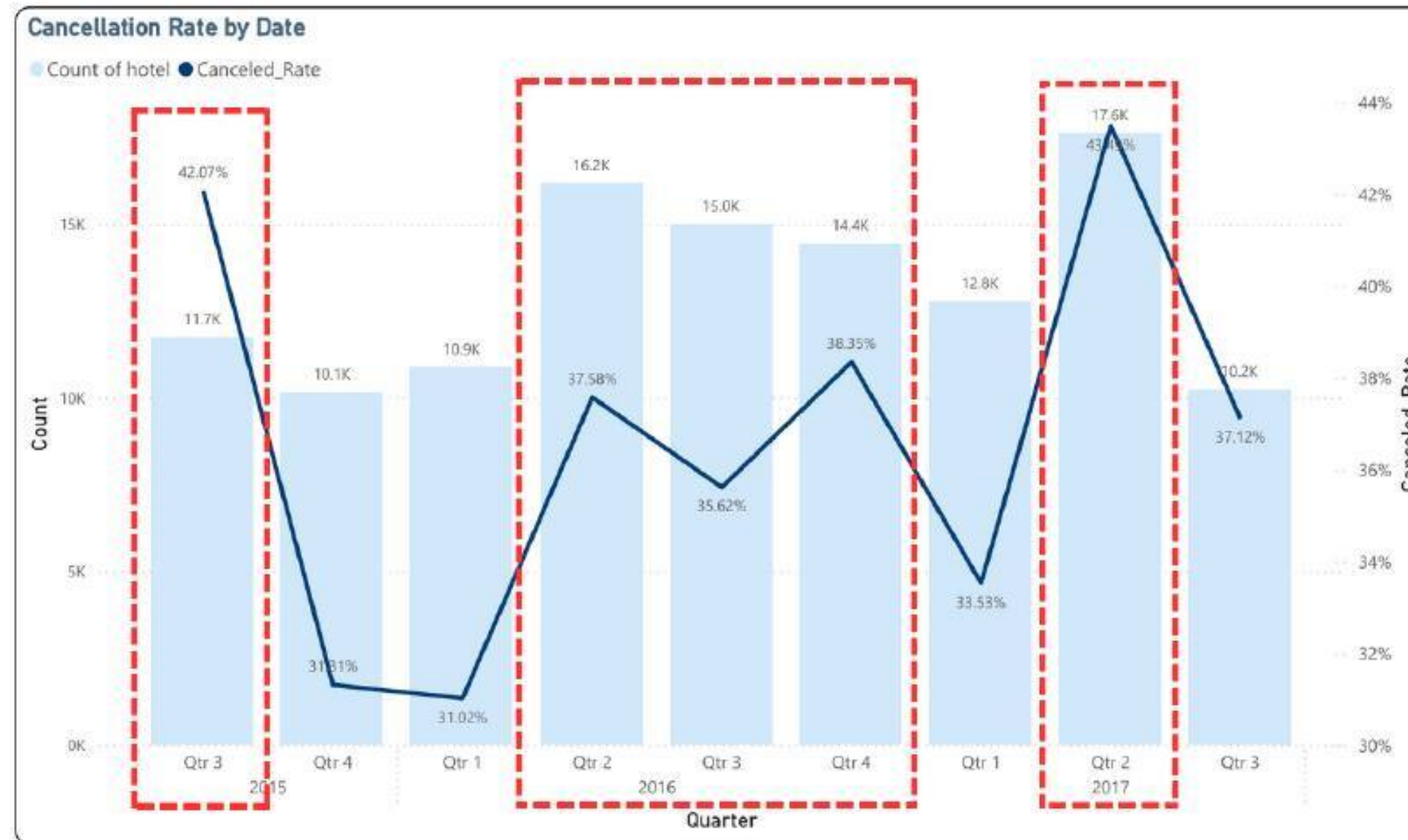
Resort Hotel



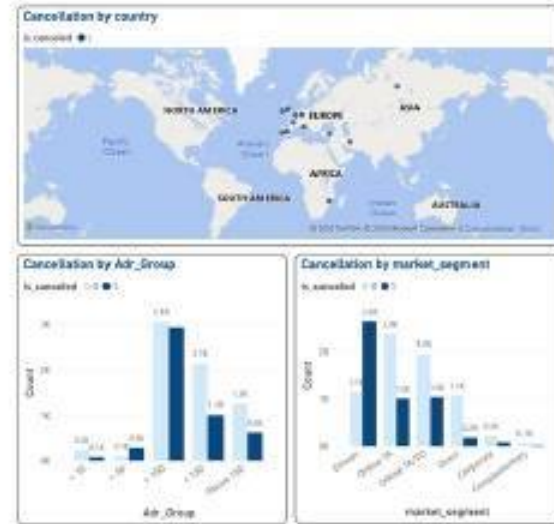
- The cancellation rate is **27.98% lower than the average**.
 - As lead time increases, the cancellation rate also tends to rise.
 - The cancellation rate fluctuates between 21.89% and 38.07%, with a trend of being **lower** at the beginning of the year (e.g., **January**) and **peaking in July** across all three years.
- Could this indicate that service quality and promotional offers are not meeting customer expectations, leading to cancellations?
- The highest cancellation rate is observed in the **"Groups" segment (42.42%)**; The **"Transient"** customer type has a cancellation rate of **31.50%**.
 - Customers with more special requests tend to have a lower cancellation rate compared to those with no special requests.



- Earlier bookings have a higher cancellation rate (bookings made more than 6 months in advance had a cancellation rate that exceeded the success rate)
- Last-minute bookings (same day or within a month) show very low cancellation rates, indicating more committed guests.

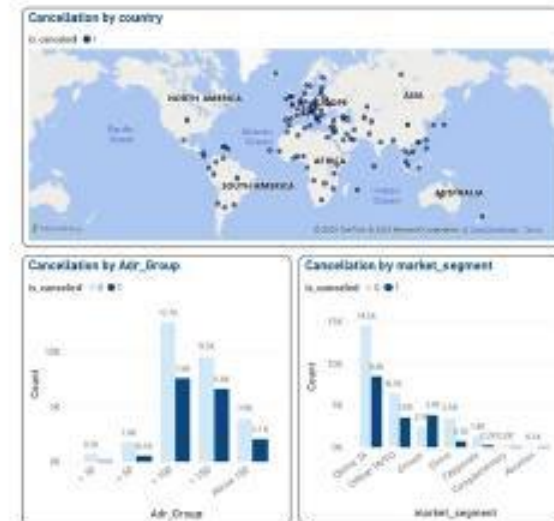


- The cancellation rate in all quarters was **over 30%**, which is **relatively high**.
- Peak seasons often come with high cancellation rates (The **highest** was in **Q2 2017: 43,4%**).
- **Booking volume** does **not always correlate** with **cancellation rate** (e.g., Q3 2015).
- The cancellation rate **fluctuates significantly over time**.



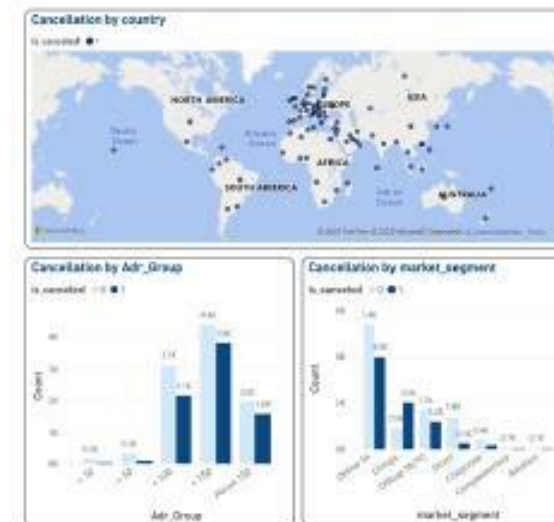
Q3/2015

- Prices (Adr) were predominantly in the 50 to 100 range
- More than half of the cancellations** came from **Groups** bookings; the rest were mostly from Online TA and Offline TA/TO
- In Q3 2015, the European migration crisis may have affected tourists' entry into the region. Additionally, the unstable situation could have discouraged tourists from traveling and led to booking cancellations



Q2-Q4/2016

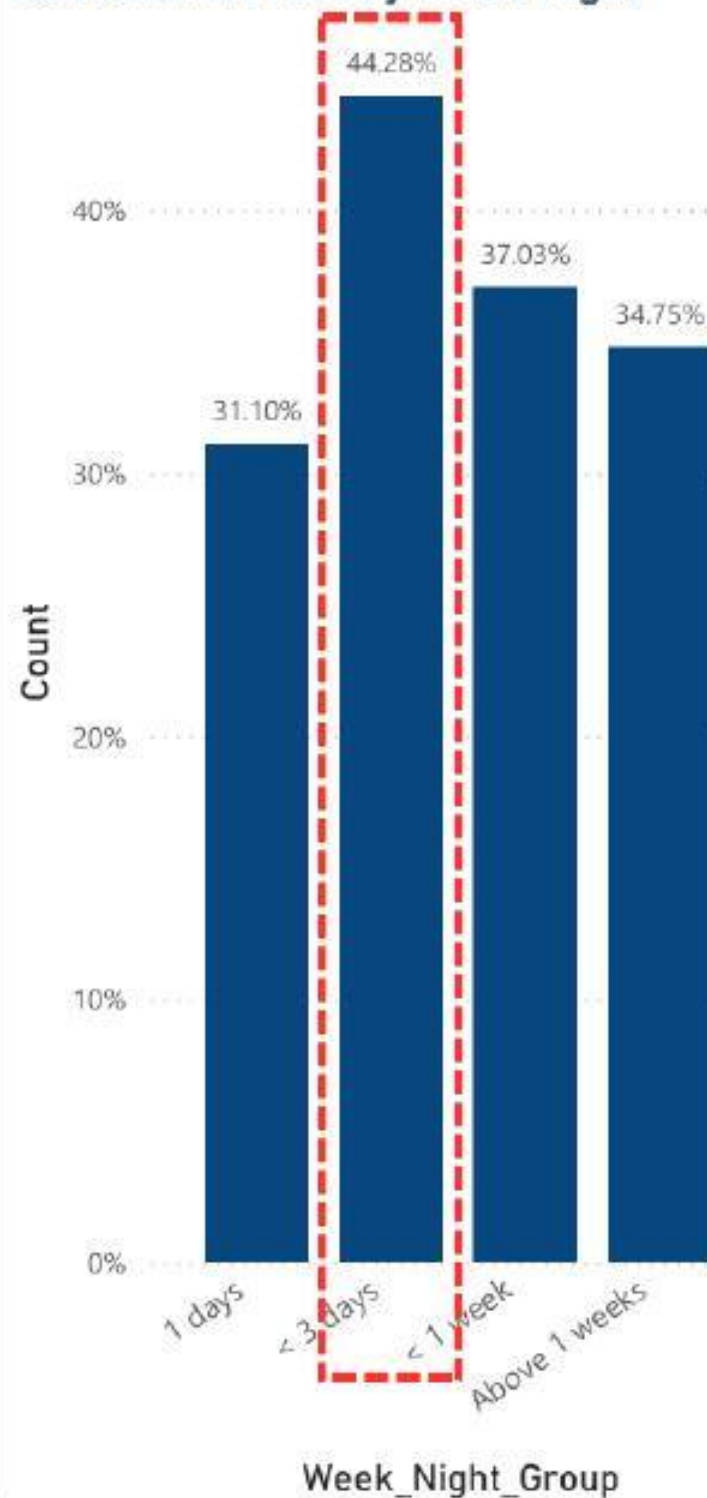
- Prices (Adr) were predominantly in the 50 to 150 range
- Cancellations were mainly from market segments: the **highest rate from Online TA (almost 19%)**, followed by Groups and Offline TA/TO
- From Q2 to Q4 2016, the UEFA Euro 2016 tournament in France, along with political events such as terrorist attacks across Europe and the U.S. presidential election, may have disrupted travel plans and contributed to an increase in booking cancellations among tourists.



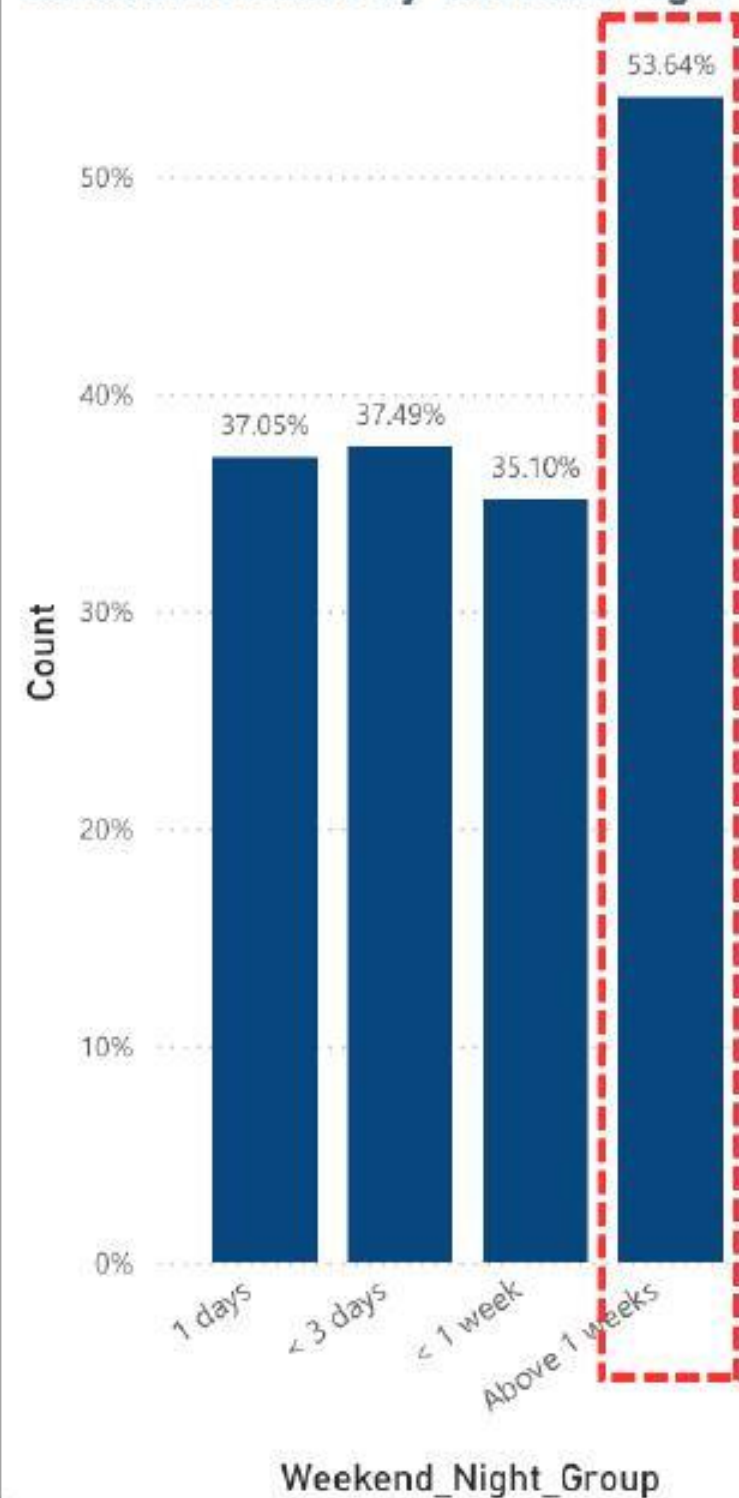
Q2/2017

- Prices (Adr) were predominantly in the 100 to 150 range
- Cancellations were mainly from market segments: the **highest rate from Online TA (almost 23%)**, followed by Groups and Offline TA/TO

Cancellation rate by Week Night



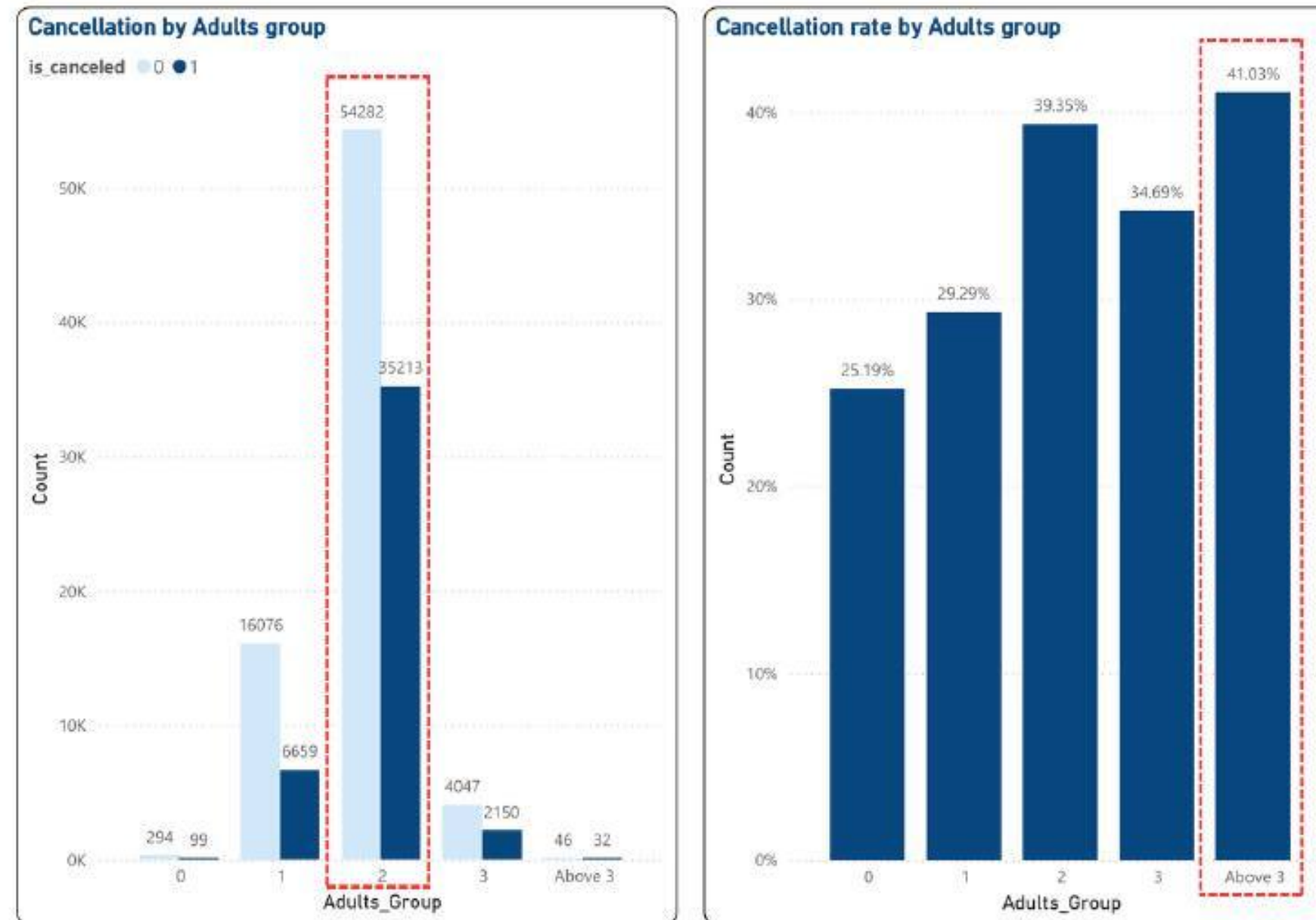
Cancellation Rate by Weekend Night



Cancellation by Week and Weekend Night

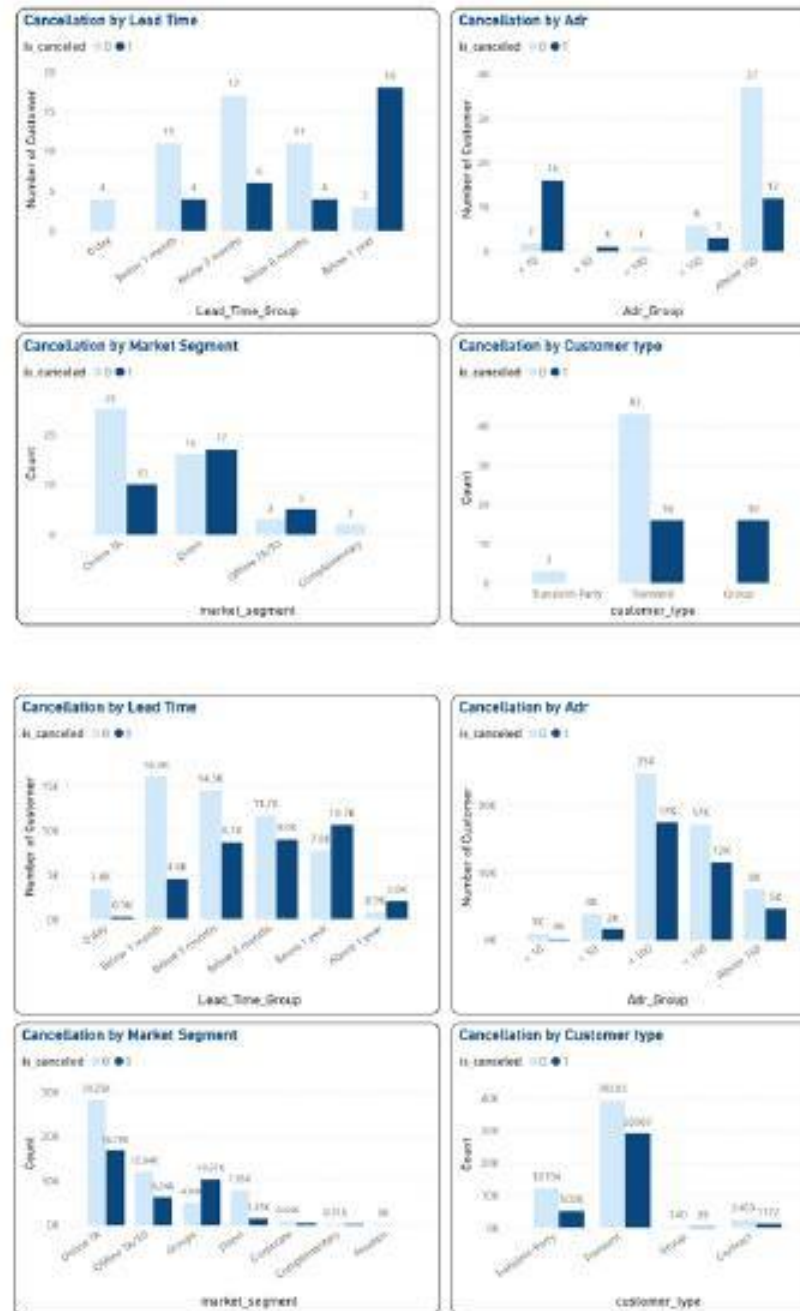


- **Guests booking over 1 week at weekends:** cancellation rate up to 53.64%.
- **Guests booking 2–3 weeknights:** 44.28% — also notable.
- **Cancellation rates tend to rise during peak seasons:** April and July (tourist months) see the highest bookings and cancellations → risk of fake/placeholder bookings.
- **Weekday bookings dominate over weekends:** Weeknight stays are 2–3 times higher weekend night



- **Bookings with two adults** were the **most common** and also had a **high cancellation rate (39.35%)**.
- The **highest cancellation rate** was observed in the group with **more than three adults (41.03%)**, although this group had the lowest number of bookings.

Zoom in
The group of canceled bookings shared the following characteristics:

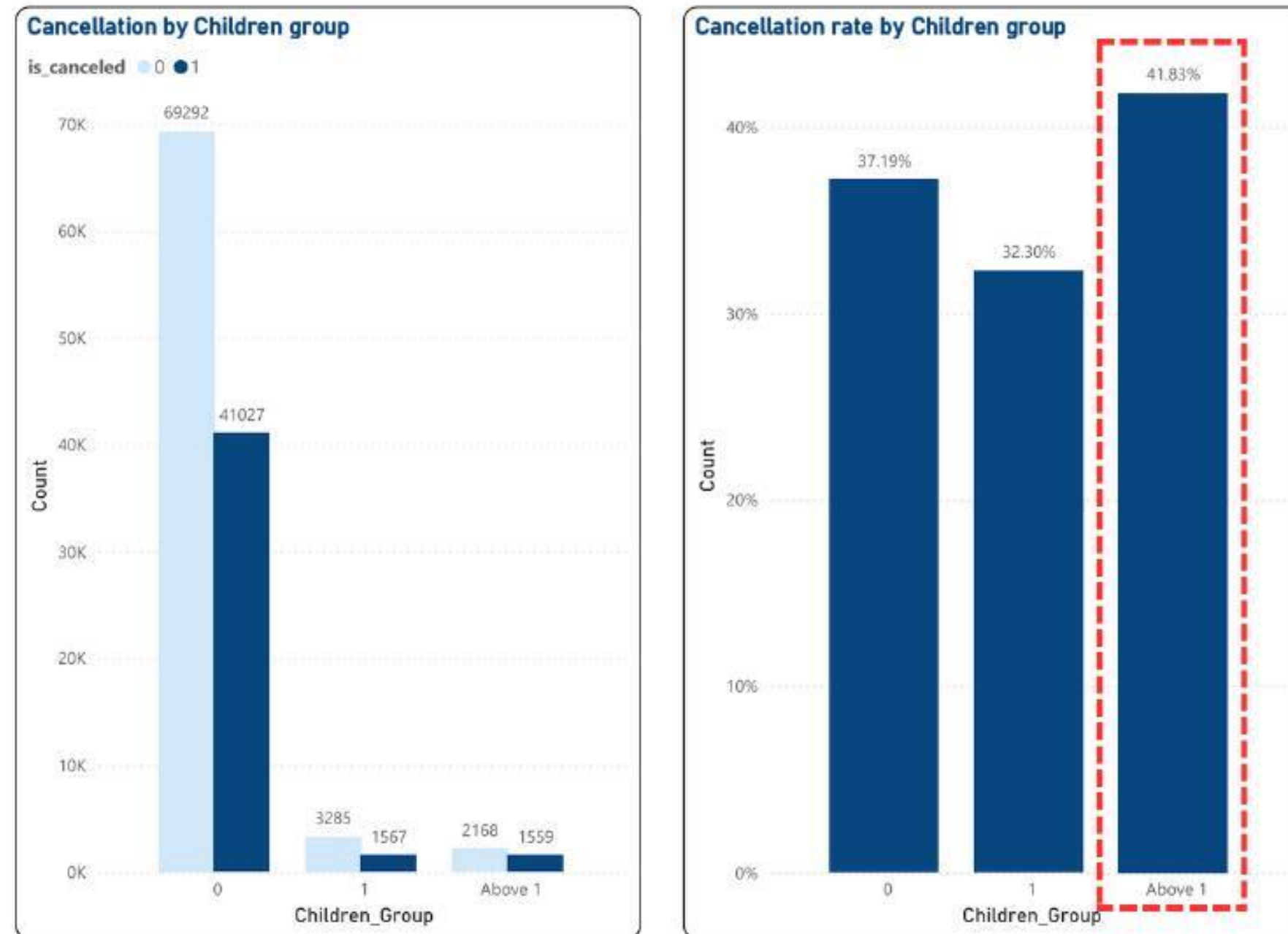


“Adults above 3”

- Mainly from lead time of 1 to 6 months; lead times **over 6 months** had a **higher cancellation rate than success rate**.
- Most cancellations were from bookings with an **ADR above 150**.
- **Online TA** accounted for the highest number and rate of cancellations (91% of total cancellations).
- Primarily from **Transient** customer type (~95% of total cancellations).

“Adults = 2”

- Mainly due to lead times over 1 month, especially **over 6 months** where **cancellations exceeded successful stays**.
- Cancellations were mostly associated with bookings **priced above 50**.
- High cancellation rates observed in **Online TA, Offline TA/TO, and Groups** segments (Groups had a higher cancellation rate than success rate).
- Primarily from **Transient** customer type (accounting for ~82% of total cancellations).

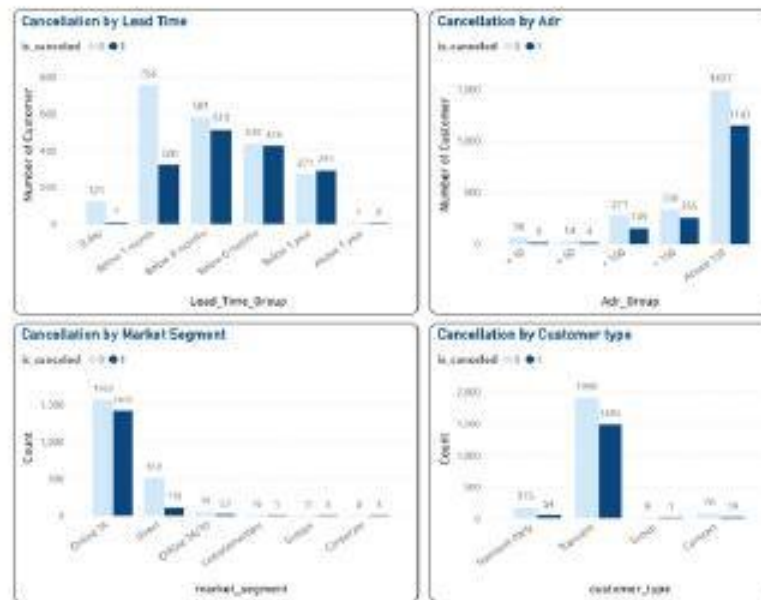


- Bookings with **more than one child** have the lowest volume but the **highest cancellation rate (41.83%)**.
- Bookings **without children** make up the majority and also have a **high cancellation rate (37.19%)**.



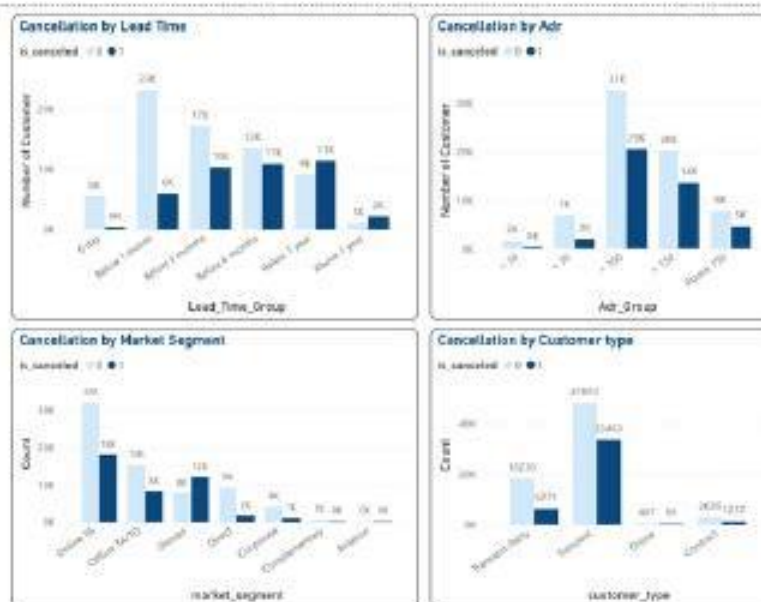
Zoom in

The group of canceled bookings shared the following characteristics:



“Children above 1”

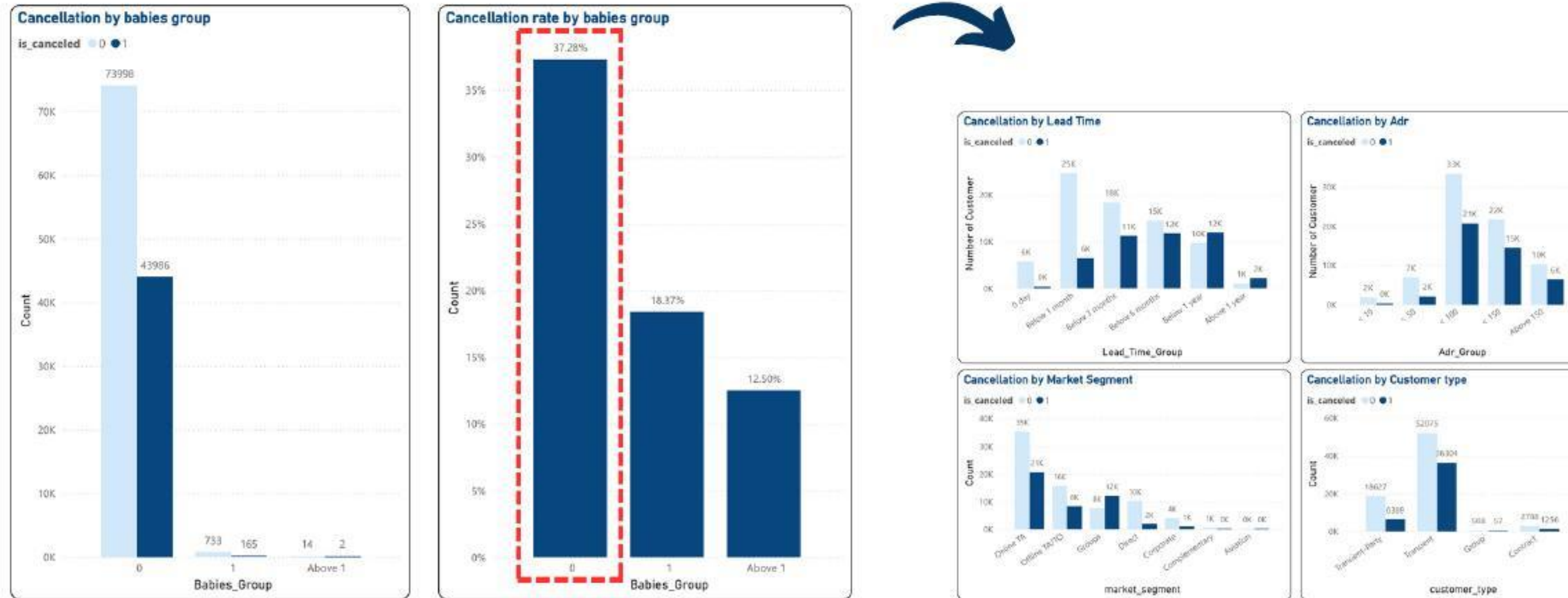
- Mainly from lead time of 1 to 6 months; lead times **over 6 months had a higher cancellation rate than success rate**.
- Most cancellations were from bookings with an **ADR above 150**.
- **Online TA** accounted for the **highest** number and rate of cancellations (**91% of total cancellations**).
- Primarily from **Transient** customer type (~**95% of total cancellations**).



“Without Children ”

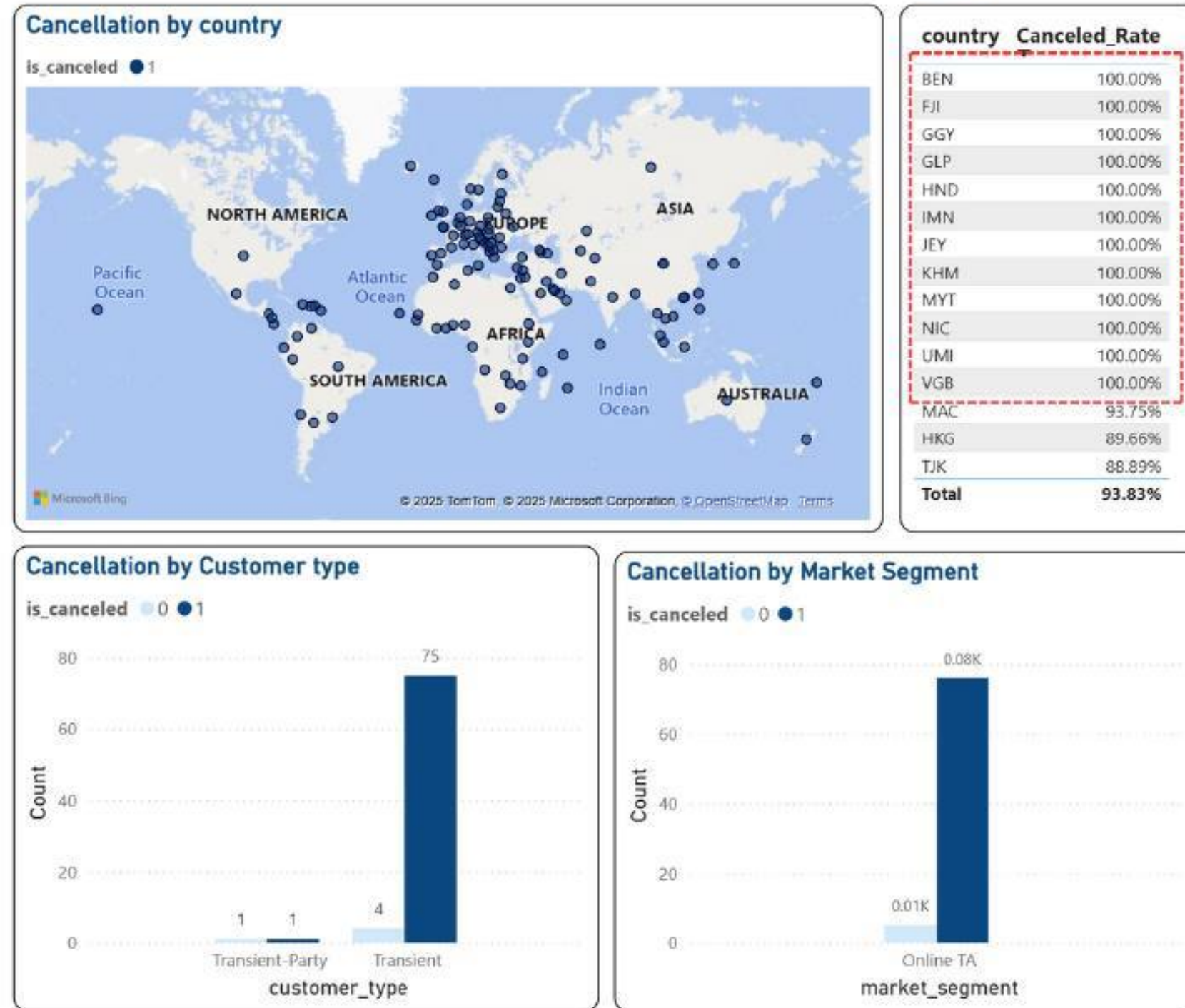
- Mainly from lead time of 1 to 6 months; **lead times over 6 months had a higher cancellation rate than success rate**.
- Most cancellations were from bookings with an **ADR above 50**
- High cancellation rates observed in **Online TA, Offline TA/TO, and Groups** segments (Groups had a higher cancellation rate than success rate)
- Primarily from **Transient** customer type (~**82% of total cancellations**).

Zoom in booking without Baby

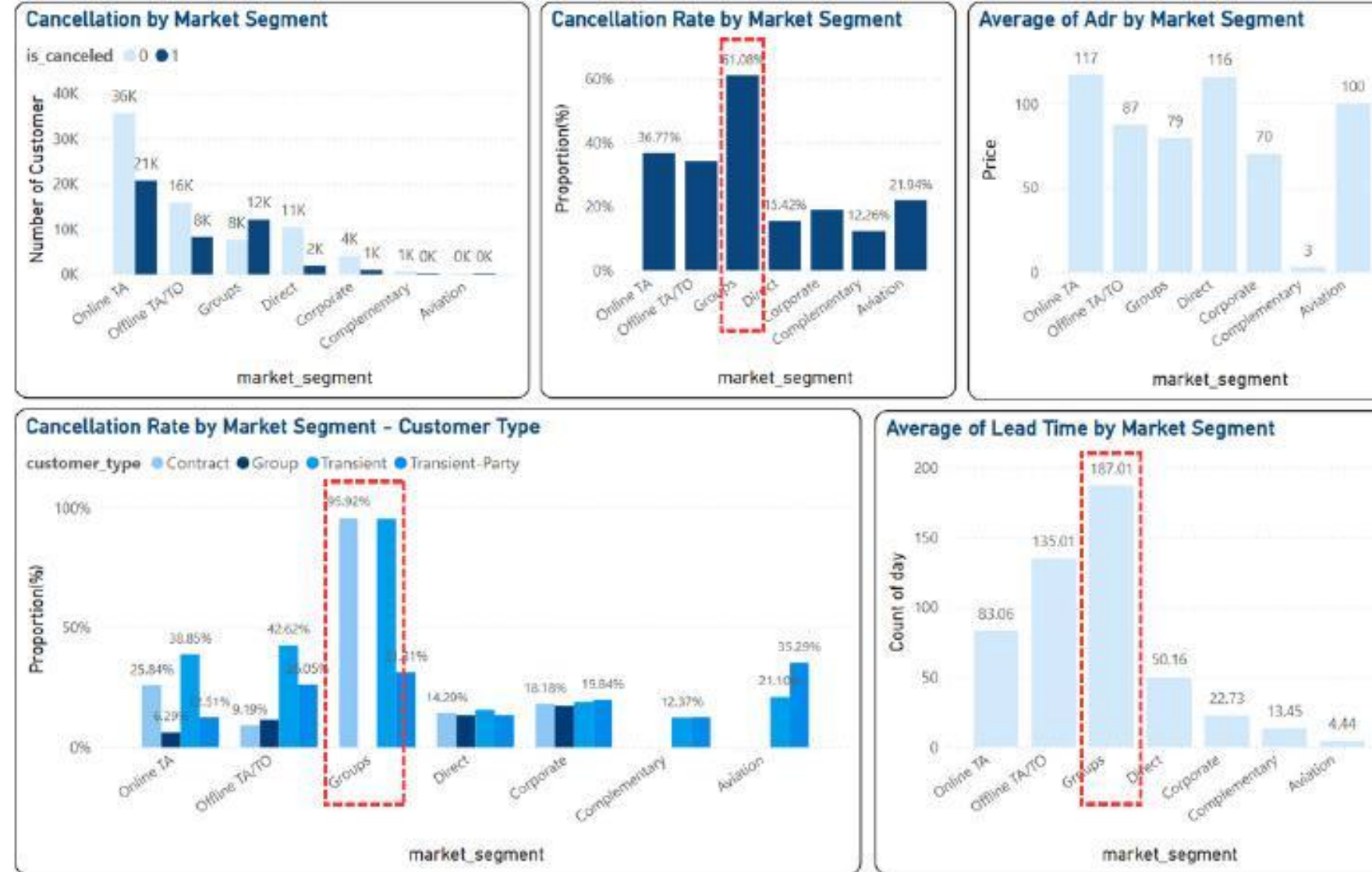


- Bookings **without babies** account for the **vast majority**, with a cancellation rate of **37.28%**.
- Cancellation rate drops significantly when babies are present—only 12.5% with more than one baby.
- A deeper analysis of bookings without children shows that :
 - Most bookings and cancellations were concentrated in ADR above 50
 - High cancellation rates observed in **Online TA, Offline TA/TO, and Groups** segments (Groups had a higher cancellation rate than success rate)
 - Primarily from **Transient** customer type (~83% of total cancellations).

Zoom in on the top 15 countries with the highest cancellation rates.



- Cancellations were widely ***distributed across regions***.
- Many countries had ***extremely high cancellation rates***, especially those with a 100% cancellation rate such as BEN, FJI, GGY, GLP, HND, etc.
- When zooming in on the top 15 countries with the highest cancellation rates, it is evident that all bookings were made through ***Online TA (100%)***, and ***almost all were from Transient customer types***.

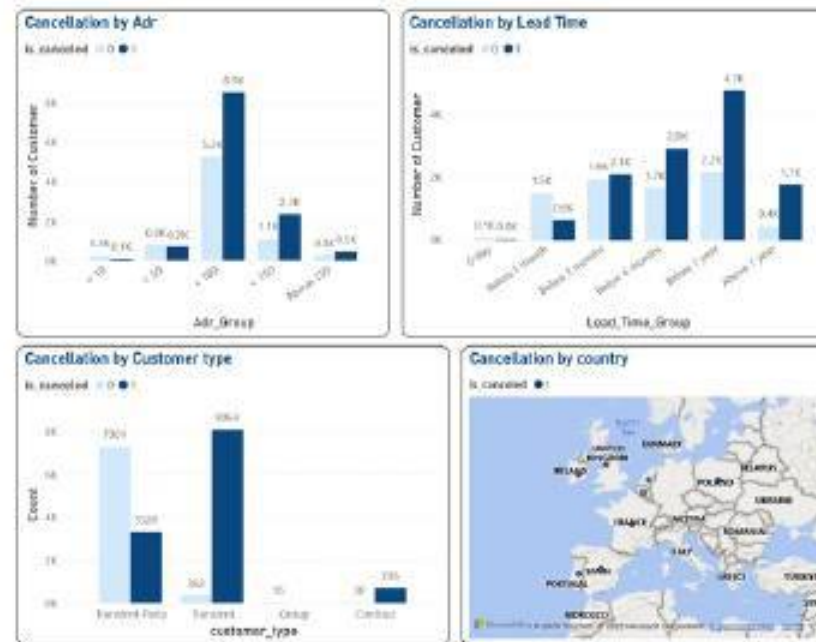


- **Online TA** accounted for the **highest number of cancellations**, with 21,000 bookings.
- The “**Groups**” segment had the **highest cancellation rate at 61.08%**, despite not having a large number of guests. This was especially true for customers with **Contract and Transient types**, both showing cancellation rates **above 95%**. The **average lead time was also high** (187 days), suggesting that bookings were made well in advance and may have been canceled due to changes in travel plans.
- **ADR was not directly proportional to the cancellation rate.** For example, Direct bookings had one of the highest ADRs (around 116) but a low cancellation rate, while Group bookings had a lower ADR but the highest cancellation rate.



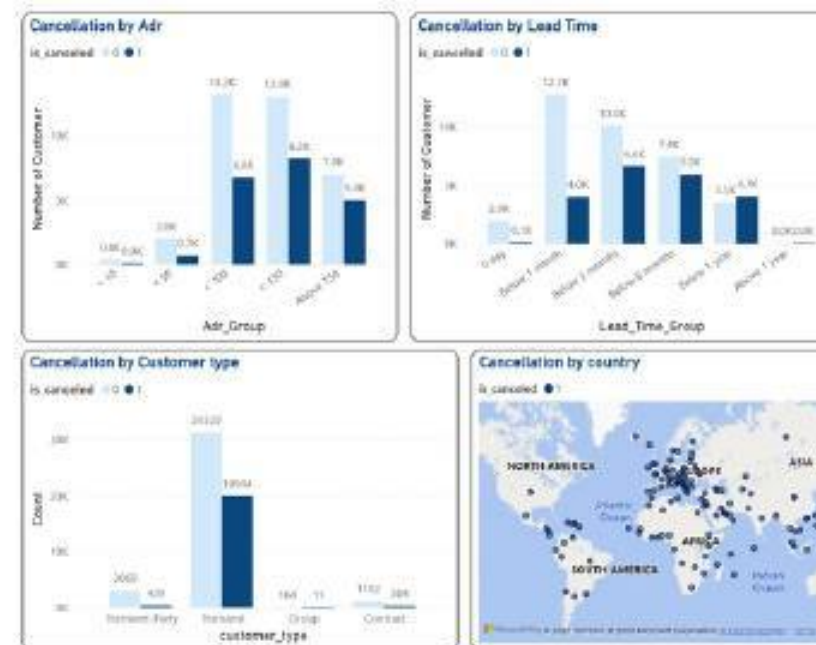
Zoom in

The group of canceled bookings shared the following characteristics:



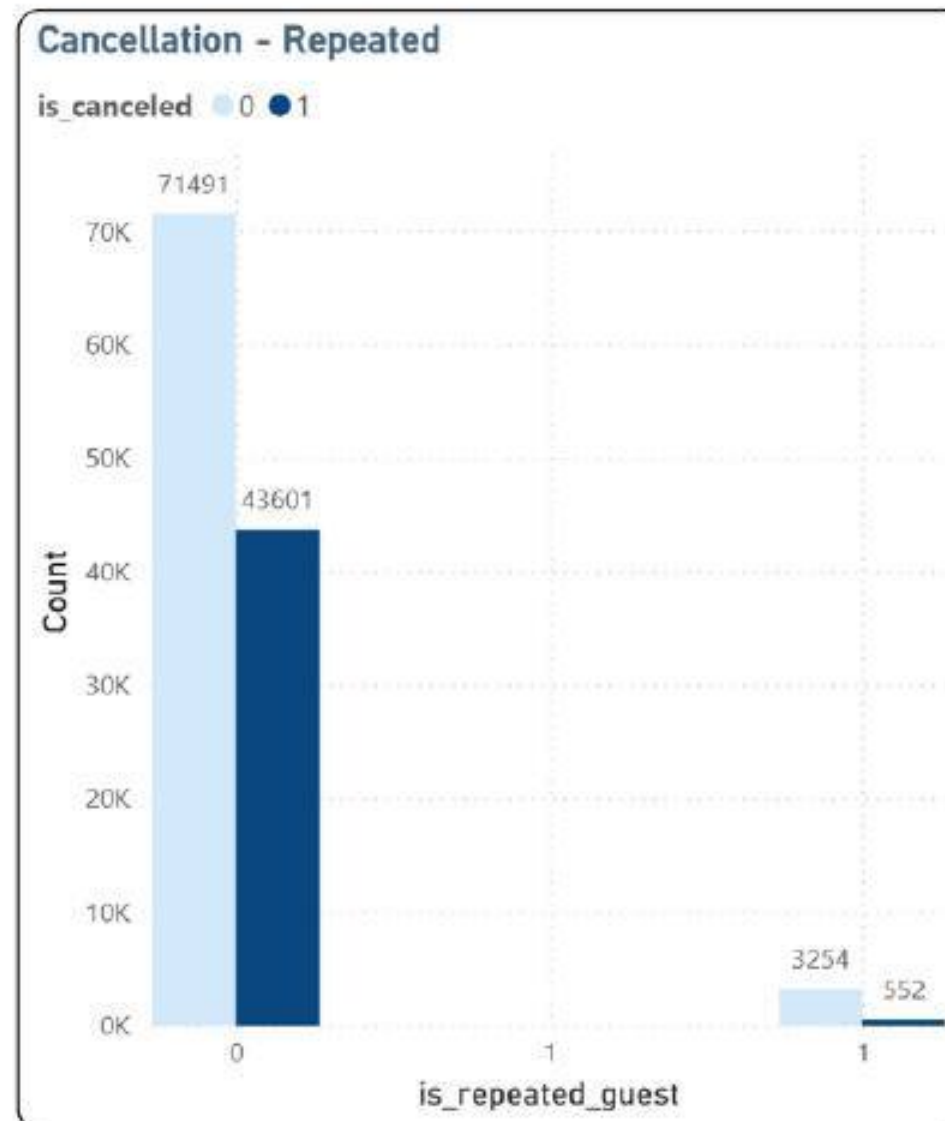
“Groups”

- ADR was mostly in the range of 50 to 100.
- Bookings with a lead time of more than 3 months had a relatively high cancellation rate; the **longer the lead time**, the **higher the cancellation rate**.
- **Most cancellations** came from the **Transient** customer type, accounting for **nearly 41%** of all cancellations.
- Cancellations were mainly from **Western European** countries, such as the United Kingdom, France, and Spain.



“Online TA”

- ADR was above 50.
- Bookings with a lead time of over one month had a high number of cancellations, especially those with a **lead time below one year**, where the **cancellation rate was even higher than the success rate**.
- Most cancellations came from the **Transient** customer type, accounting for **nearly 35%** of all cancellations.
- Cancellations were widely distributed across regions.



- The **repeated guests** was very **low (3.2%)**. Among them, 552 canceled their bookings, with **a lower cancellation rate (~14.5%)** compared to the overall rate of 37.14%.
→ **Repeated guests tend to cancel less.**

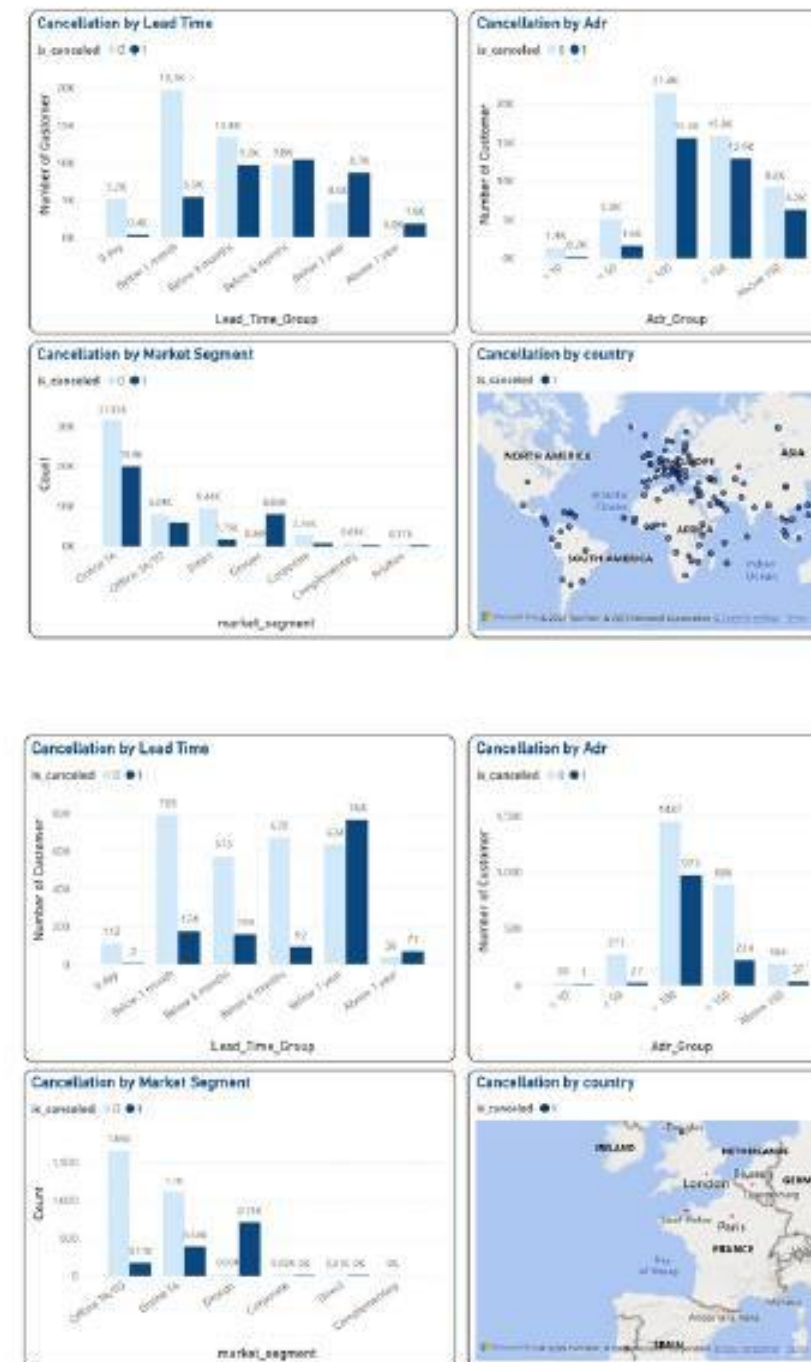


Compared to all canceled bookings in the dataset, cancellations by Repeated Guests show some distinct patterns:

- The highest cancellation rate was in the **lead time group of 6 months to 1 year**. Unlike new customers in the full dataset, Repeated Guests did not book more than 1 year in advance → This suggests that early bookings may lead to more cancellations due to changes in plans.
- The cancellation rate gradually **declined and stabilized, dropping from 82.7% to under 10%**, while the overall dataset showed fluctuating trends → This indicates that Repeated Guests improved their booking behavior over time and maintained consistency.
- Cancellations were **highest in the "Groups" segment**, followed by Offline TA/TO.
- Bookings with **0 special requests** had the **highest cancellation rate**, but the number of **special requests** did **not correlate** with cancellation rates as it did in the full dataset.
- The **Transient-Party** customer type had the **highest cancellation rate** (unlike the full dataset, where Transient had the highest), followed by Contract.



Zoom in



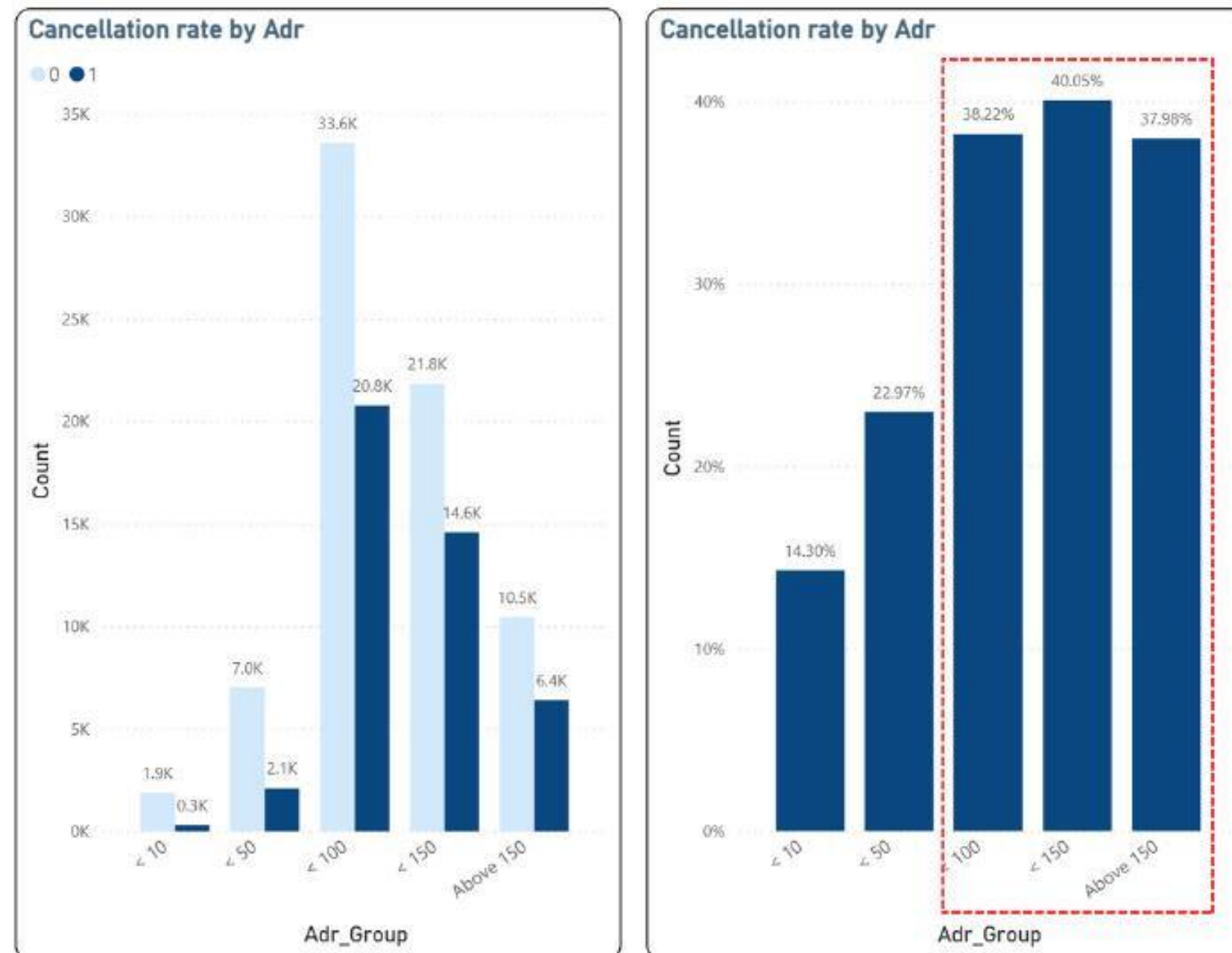
- The **Transient** customer type had the highest number of bookings and the **highest cancellation rate (40.89%)**.
- Next was the **Contract** customer type, which had a **lower number of bookings** but a relatively **high cancellation rate (30.96%)**.

“Transient”

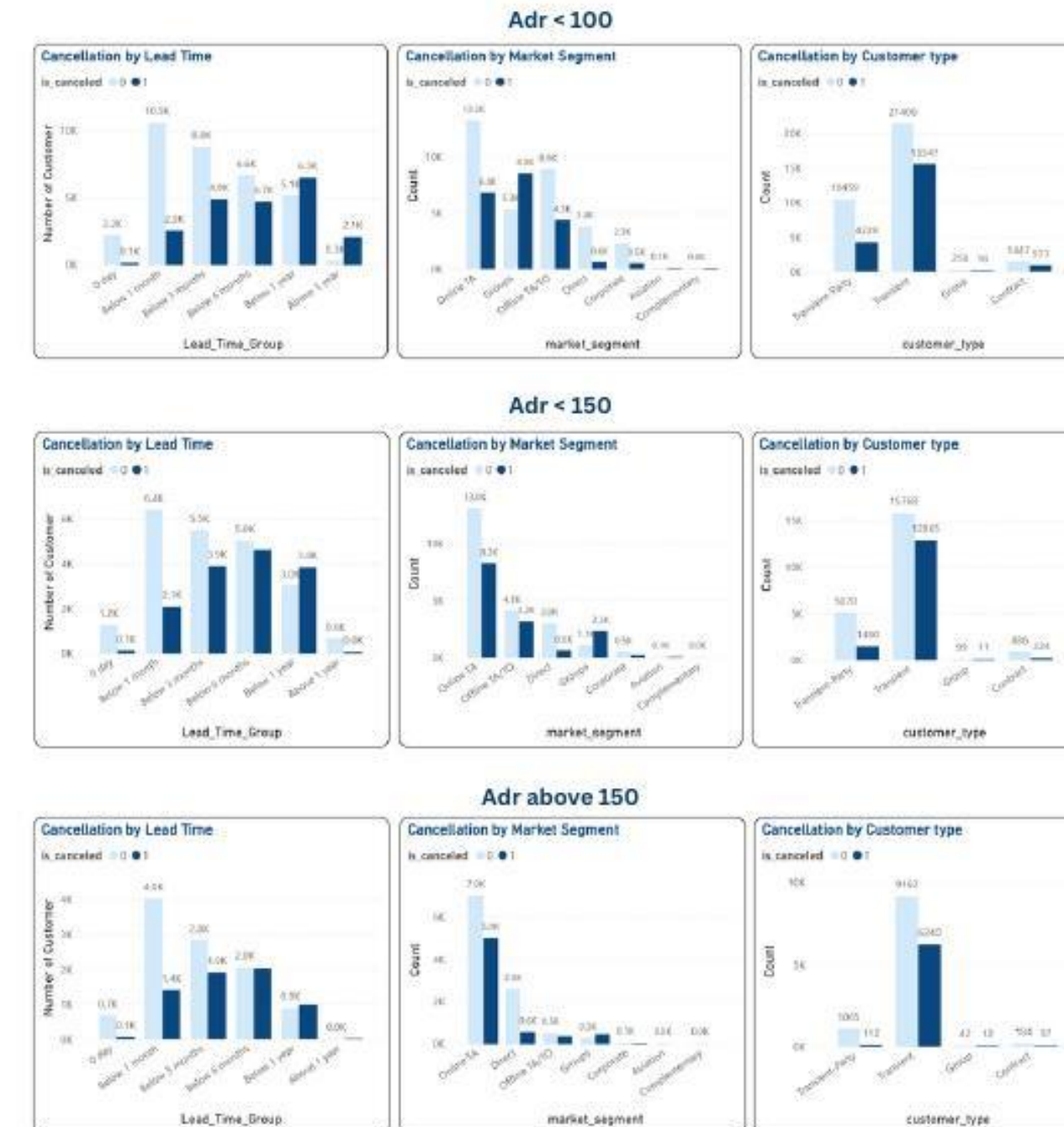
- Bookings with a lead time from 3 months to 1 year had a higher cancellation rate than success.
- ADR was mostly above 50.
- The **largest** number of bookings and cancellations came from the **“Online TA”** market segment, while the **highest cancellation rate** was observed in the **“Groups”** segment, despite its lower volume.
- Cancellations were widely distributed across regions.

“Contract”

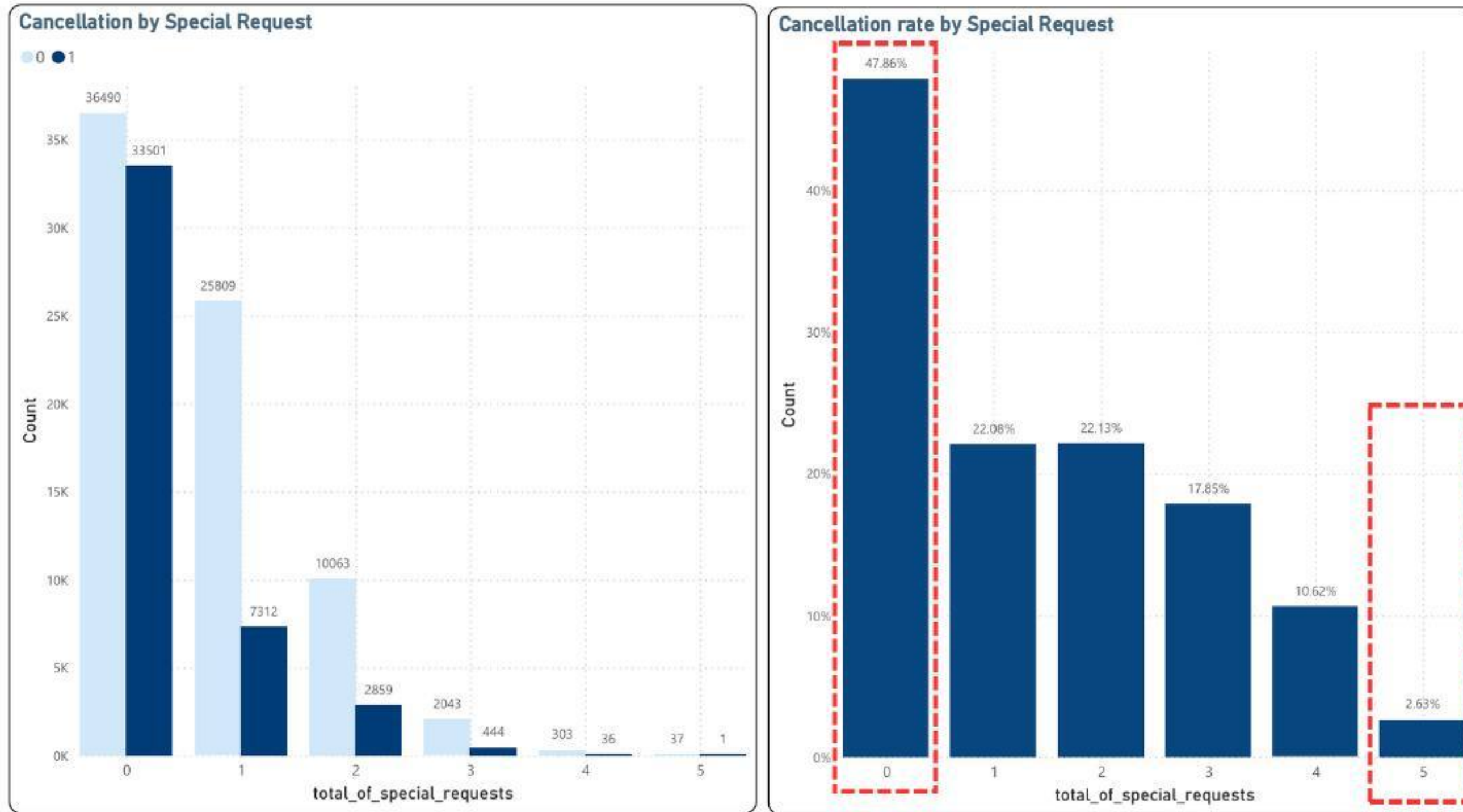
- Bookings with a **lead time from 6 months to 1 year** had the highest number of cancellations, with the cancellation rate exceeding the success rate.
- ADR was mainly in the range of 50 to 100.
- Most booking cancellations** belonged to the **“Groups”** segment, accounting for **over 95%** of that segment’s total bookings.
- Cancellations were primarily concentrated in **Western Europe**, such as the United Kingdom, France, and Spain.



- **Adr above 50 had a high cancellation rate**, with the highest rate observed in **the 100–150 range (40.05%)**.
- Lower price ranges have lower cancellation rates, with the **lowest at Adr < 10 (14.3%)**. Maybe it's the low price or the attractive promotion.



- **The longer the Lead Time, the higher the cancellation rate.**
- **Online TA** accounts for the **largest** number and rate of cancellations, but **Groups** segment has the **highest cancellation rate**, with cancellation rates exceeding success rates
- **Transient** Customer Type accounted for the **highest number and rate of cancellations**



- Bookings with no special requests have the highest volume and cancellation rate (47.86%).
- Both booking volume and cancellation rate decrease as the number of special requests increases — only 2.63% cancel when there are 5 special requests.

Through the Exploratory Data Analysis (EDA) process, we identified several **key factors influencing** the hotel booking cancellation rate. Notably, customers in the **Transient** customer type, bookings through the **Online TA and Groups** segments showed both **high volumes and high cancellation rates**. Additionally, we observed that customers with more special requests tended to have a lower cancellation rate, indicating stronger intent to stay.

However, descriptive analysis alone is not sufficient to predict customer behavior with high accuracy. This is where Machine Learning (ML) becomes essential — enabling us to learn patterns from data, make accurate predictions, and automate decision-making processes.

Given the dataset's scale — 118,898 rows and 17 columns — and the goal of identifying customers who are likely to cancel their bookings (so that proactive retention strategies can be implemented), we selected the **Random Forest** model. This algorithm was chosen for its ability to:

- **Handle complex and high-dimensional data effectively**
- **Improve accuracy by aggregating multiple decision trees**
- **Reduce overfitting through result averaging**

By building on the foundation laid by EDA, **Machine Learning enables more effective strategies to minimize cancellations and improve booking retention.**

Train test split

```
# X, Y
X = df.drop(columns=['is_canceled'])
y = df['is_canceled']
```

```
# train test split
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1, stratify=y)
```

Data Preprocessing

```
# One-Hot Encoding
X_train = pd.get_dummies(X_train, columns=['hotel', 'market_segment', 'customer_type'])
X_test = pd.get_dummies(X_test, columns=['hotel', 'market_segment', 'customer_type'])
```

```
#LabelEncoder
le = LabelEncoder()
X_train['country'] = le.fit_transform(X_train['country'])
X_test['country'] = le.fit_transform(X_test['country'])
```

```
# Convert all True/False values to 1/0
X_train = X_train.astype(int)
X_test = X_test.astype(int)
```

Optimizing Hyperparameters with Grid Search

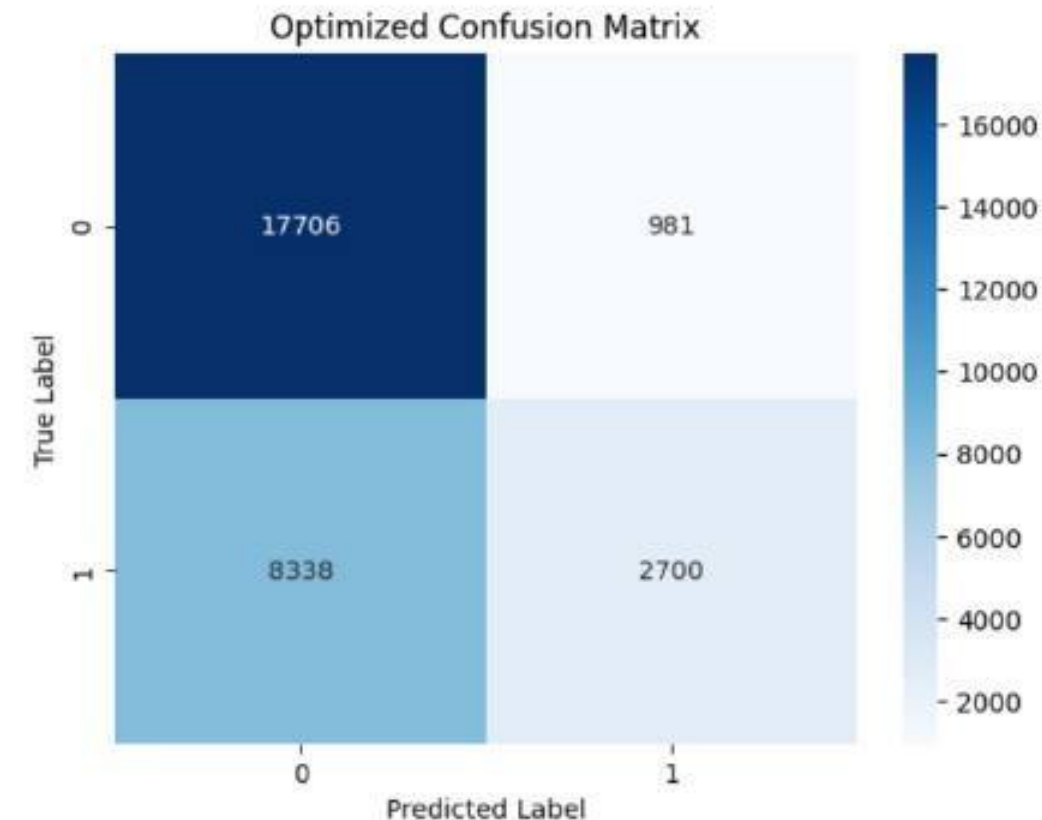
```
#hyperparameters turning
#from sklearn.model_selection import RandomizedSearchCV
rf = RandomForestClassifier()
from sklearn.model_selection import GridSearchCV
grid_space = {'max_depth':[3, 5, 10],
              'n_estimators':[100, 200, 300],
              'criterion':['gini', 'entropy'],
              }
rf_grid = GridSearchCV(rf, param_grid=grid_space, scoring='accuracy')
# Huấn luyện Grid Search
rf_grid.fit(X_train, y_train)

print("Best Parameters:", rf_grid.best_params_)

Best Parameters: {'criterion': 'gini', 'max_depth': 10, 'n_estimators': 200}
```

Optimized Accuracy Score: 0.6864928511354079

	precision	recall	f1-score	support
0	0.68	0.95	0.79	18687
1	0.73	0.24	0.37	11038
accuracy			0.69	29725
macro avg	0.71	0.60	0.58	29725
weighted avg	0.70	0.69	0.63	29725



- Accuracy is 69%, the model is heavily biased toward class 0 (**data imbalance**).
- Recall for class 1 is low (24%) → many class 1 cases are missed.
- F1-score for class 1 is also low → the model performs poorly in identifying class 1.

```
print(Counter(y_train))
print(Counter(y_test))
```

```
Counter({0: 56058, 1: 33115})
Counter({0: 18687, 1: 11038})
```

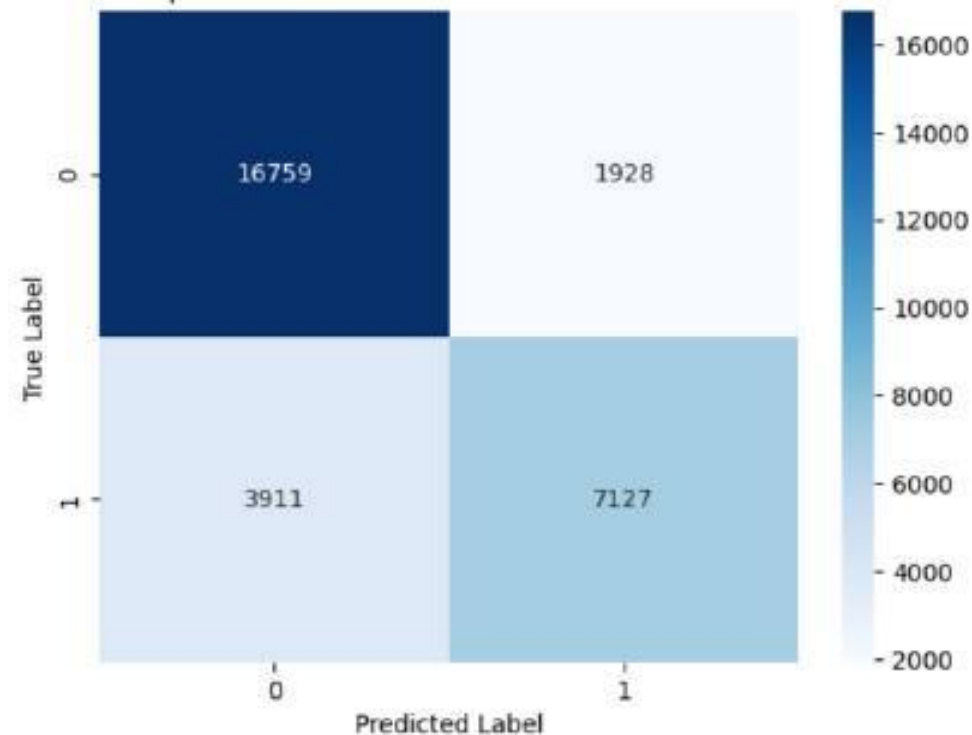
```
from imblearn.over_sampling import SMOTE
```

```
smote = SMOTE(random_state=42)
X_train_res, y_train_res = smote.fit_resample(X_train, y_train)
Counter(y_train_res)
```

```
Counter({0: 56058, 1: 56058})
```

	precision	recall	f1-score	support
0	0.81	0.90	0.85	18687
1	0.79	0.65	0.71	11038
accuracy			0.80	29725
macro avg	0.80	0.77	0.78	29725
weighted avg	0.80	0.80	0.80	29725

Optimized Confusion Matrix- After SMOTE



Class 1 is significantly smaller than Class 0, causing an imbalance.

=> Apply SMOTE to address this by increasing the smaller class to match the larger one.

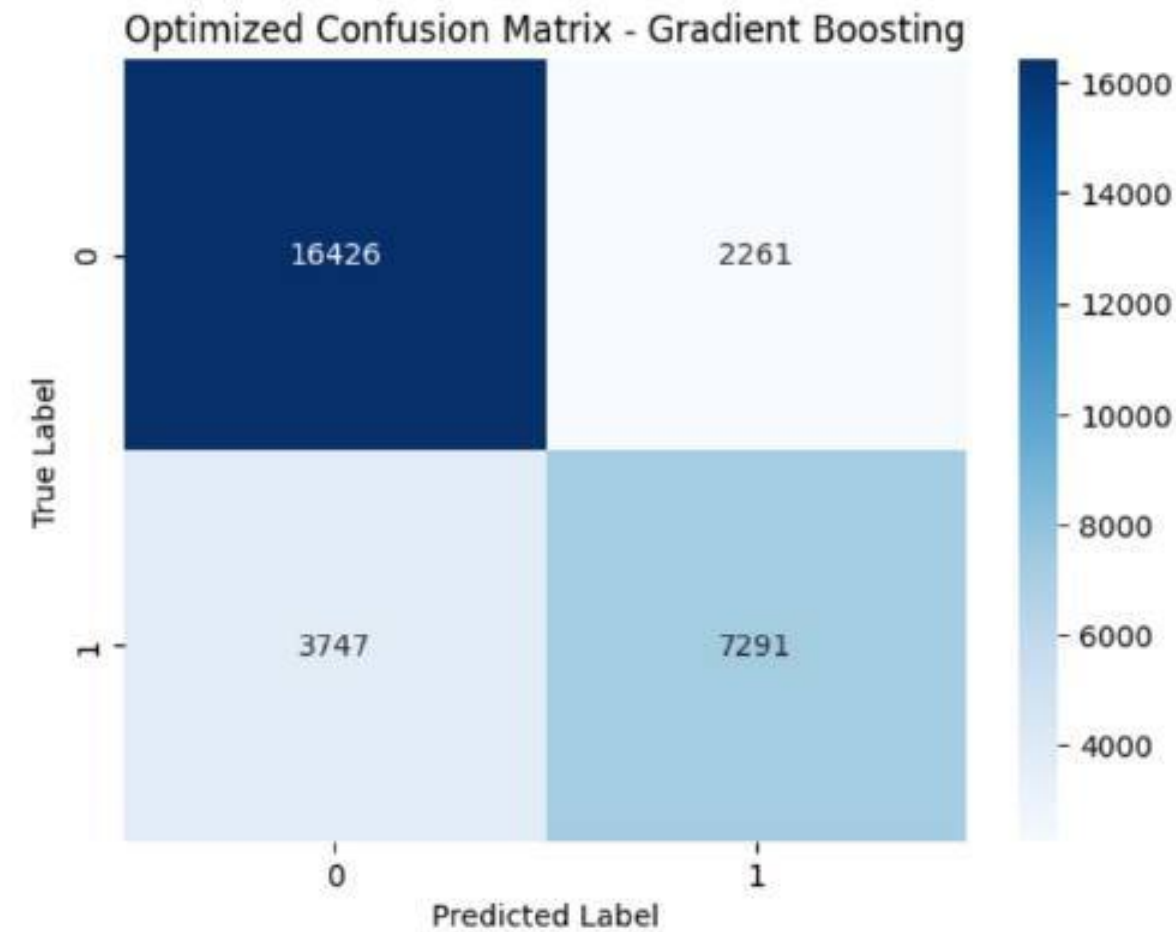
- Using SMOTE (oversampling class 1) significantly improves model performance for class 1.
- Recall (65%) and F1-score (71%) for class 1 increase noticeably.
- Accuracy also improves from 69% to 80%.

→ Since the goal is to reduce booking cancellations, the model with SMOTE is more suitable.

→ Other techniques, such as Gradient Boosting, can also be explored to false predictions.

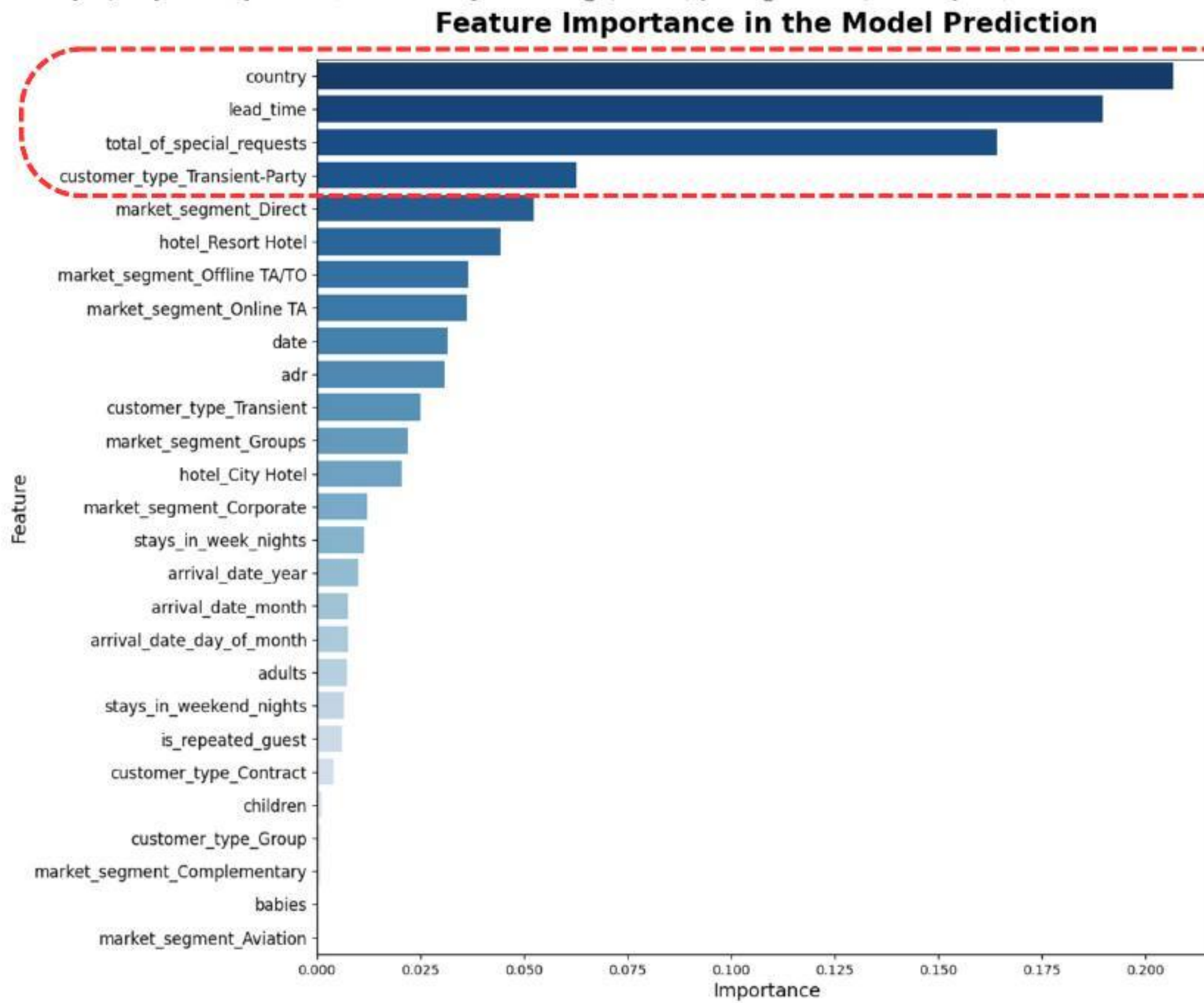
Optimized Accuracy Score: 0.7978805719091674

	precision	recall	f1-score	support
0	0.81	0.88	0.85	18687
1	0.76	0.66	0.71	11038
accuracy			0.80	29725
macro avg	0.79	0.77	0.78	29725
weighted avg	0.80	0.80	0.79	29725



- Accuracy: 80% (same as the model with SMOTE)
- Recall (Class 1): 66% — slightly higher than the model with SMOTE (64%)

=> Since the goal is to reduce booking cancellations, Gradient Boosting performs slightly better than the model with SMOTE



- **Country:** The most important factor influencing cancellations
- **Lead_time:** The earlier the booking, the higher the risk of cancellation
- **Total_of_special_requests:** Cancellation rate decreases as the number of special requests increases
- **Customer_type_Transient-Party:** Short-term group travelers have high cancellation risk
- The remaining factors have gradually less impact on the cancellation rate.

Booking Behavior & Stay Pattern

- Lead time → higher cancellation risk.
- More special requests → lower cancellation → more serious guests.
- Weekday stays slightly relate to cancellations
- Weekend stays less so, but retention is higher for leisure travelers.

Customer Profile & Type

- Country is the strongest factor, *Western European* countries should be given *special attention* → monitor by market.
- Repeated guests cancel less → high-value group.
- Customer type: Transient customers and Contract cancel most
- Prioritize loyal and committed segments.

Hotel, Channel & Pricing

- City Hotels have higher cancellation rates than Resort Hotels
- Market segment:
 - Online TA and Groups → higher risk.
 - Direct and Complementary → more stable.
- Promote Direct Booking for better control.
- Higher-paying guests with special requests are more likely to keep bookings.

Predictive Models

- Always handle data imbalance (e.g., SMOTE) before training the model.
- Combining resampling techniques (e.g., oversampling) with strong models like Gradient Boosting yields the best results.

Apply stricter cancellation policies for early bookings (over 6 months in advance)

- Require non-refundable deposits or enforce no-cancellation policies close to check-in dates.

Prioritize and encourage guests with multiple special requests

- Provide a more detailed request form to assess booking commitment levels.

Focus on high-risk segments
(Customer Types: Transient and Contract; Market Segments: Groups, Online TA, and Offline TA/TO)

- Send frequent reminders and re-confirm bookings closer to the stay date.
- Monitor countries with high cancellation rates:
- Implement mandatory deposits or require advance payment.

Design a loyalty program

- Provide exclusive benefits for returning guests (repeated guests)

Encourage direct bookings

- Offer incentives such as complimentary breakfast or free room upgrades for direct reservations.

Customize policies for each customer segment

- Closely monitor bookings made through Online TA and the Groups market segment, especially for Transient customer types using these channels.
- For Contract customers, maintain more flexible policies, but closely track lead time behavior.

Introduce non-refundable packages

- Target early bookers with added benefits such as breakfast or room upgrades to increase commitment.

**Develop an intelligent
booking cancellation
prediction model**

- Test alternative models to optimize recall and avoid missing potential customers.
- Other imbalance-handling techniques: Combine SMOTE with undersampling

**Incorporate country- and
market-level analysis**

- Customize policies based on high-risk countries and market segments

**Build an integrated
booking system with risk
assessment**

- Recommend policies tailored to user behavior and booking characteristics (e.g., early booking, without special requests → require deposit)

**Differentiate pricing
strategies for business
with leisure travelers**

- Apply dynamic pricing and cancellation policies based on travel type — weekday (business) vs. weekend (leisure) — to maximize revenue and minimize cancellations



Thank you
for listening!