Hospitality Cancellation Rate Prediction

Kitae Kim, Reina Chen, Sally Lee

Meet The B4 TEAM!

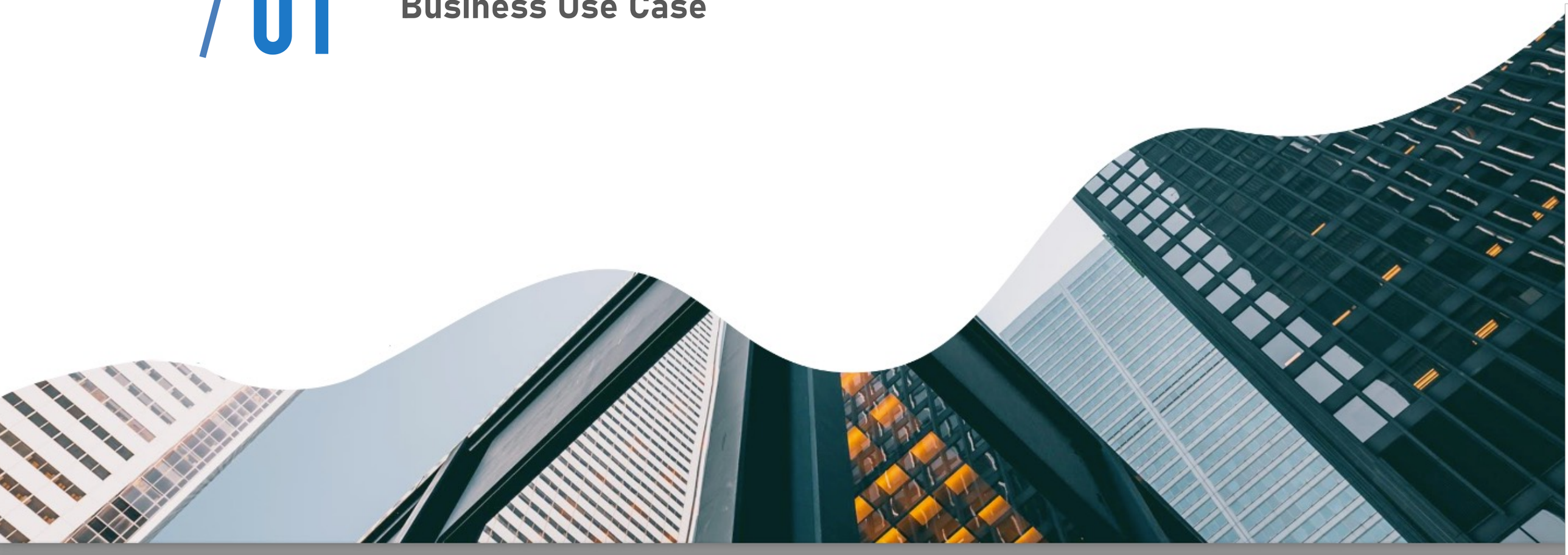
Reina Chen

Sally Lee

Kitae Kim

/ 01

Executive Summary and Business Use Case

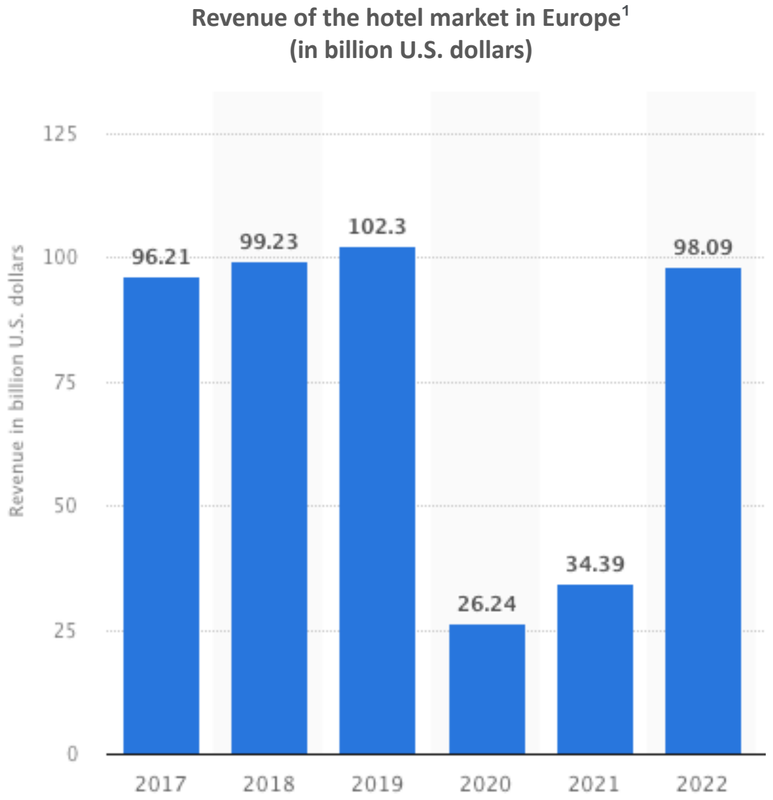


Executive Summary

The hospitality industry encountered significant challenges during the COVID-19 pandemic due to restricted governance and health conditions[1]. Furthermore, the current client demand for flexibility in cancellation has resulted in a decline in occupancy rates.

To address this, an effective strategy is to **accurately predict cancellation rates**.

In order to enhance prediction accuracy, we adopted a mixed methods approach, utilizing customer behavior, public data, and governmental data to identify trends related to cancellations, weather, and economic conditions. Aims to optimize hotel resource allocation and maximize revenue.



Research Objective

1. Investigate the impact of internal (consumer attributes) and external factors on cancellation rates
2. Utilize data analytics and modeling to predictive dynamically cancellation rates
3. Provide data driven insights and mitigate the effects on cancellation rates
    - Identify significant feature
    - Optimize marketing strategies and operation efficiency
    - Maximize the overbook rate while maintaining customer satisfaction

Data Planning

/ 02

ETL
Data Extract

Data Profile

	Data Definition	Data Source	Data Size	Rows/Column
	Resort hotel booking	ScienceDirect website Extracted from hotels' Property Management System (PMS) SQL databases	8.1 MB	(40060, 31)
	City hotel booking	ScienceDirect website Extracted from hotels' Property Management System (PMS) SQL databases	16 MB	(79330, 31)
	Economic indicator data	IMF Website	12 KB	(27, 8)
	Weather data	Visual Crossing website	49 KB	(823, 8)
	Number of tourists	Instituto Nacional de Estadística	12 KB	(1, 31)

Data Profile – Hotel Data

Goal: To optimize cancellation rate predictions and provide tailored recommendations for hotels in Europe

Characteristics:

1. Transactions in Resort and City Hotel in Lisbon with arrival date between July 1, 2015, and August 31, 2017

2. Includes booking status (i.e. ReservationStatus), payment information (i.e. DepositType), guest profile (i.e. Adults, Children), bookings detail (i.e. ArrivalDateMonth, ReservationStatusDate), types of room reserved and assigned to the guests (i.e. ReservedRoomType, AssignedRoomType), etc.

Resort Hotel Data

	IsCanceled	LeadTime	ArrivalDateYear	ArrivalDateMonth	ArrivalDateWeekNumber	ArrivalDateDayOfMonth	...	CustomerType	ADR	RequiredCarParkingSpaces	TotalOfSpecialRequests	ReservationStatus	ReservationStatusDate
0	0	342	2015	July	27	1	...	Transient	0.0	0	0	Check-Out	2015-07-01
1	0	737	2015	July	27	1	...	Transient	0.0	0	0	Check-Out	2015-07-01
2	0	7	2015	July	27	1	...	Transient	75.0	0	0	Check-Out	2015-07-02
3	0	13	2015	July	27	1	...	Transient	75.0	0	0	Check-Out	2015-07-02
4	0	14	2015	July	27	1	...	Transient	98.0	0	1	Check-Out	2015-07-03

City Hotel Data

	IsCanceled	LeadTime	ArrivalDateYear	ArrivalDateMonth	ArrivalDateWeekNumber	ArrivalDateDayOfMonth	...	CustomerType	ADR	RequiredCarParkingSpaces	TotalOfSpecialRequests	ReservationStatus	ReservationStatusDate
0	0	6	2015	July	27	1	...	Transient	0.0	0	0	Check-Out	2015-07-03
1	1	88	2015	July	27	1	...	Transient	76.5	0	1	Canceled	2015-07-01
2	1	65	2015	July	27	1	...	Transient	68.0	0	1	Canceled	2015-04-30
3	1	92	2015	July	27	1	...	Transient	76.5	0	2	Canceled	2015-06-23
4	1	100	2015	July	27	2	...	Transient	76.5	0	1	Canceled	2015-04-02

Data Profile – Economic Data

Goal: To optimize cancellation rate predictions by identifying factors related to hotel guests' financial status, such as unemployment rate and Price Index

Characteristics:

1. Economic data for the Lisbon region between July 2015 and August 2017
2. Includes Consumer Price Index, Production Index, and unemployment rate indicators presented in monthly average values

	Date	PPI	CPI	Economic Activity, Industrial Production, Manufacturing, Index	Economic Activity, Industrial Production, Index	Industrial Production, Seasonally adjusted, Index	Unemployment, Persons, Number of	Labor Markets, Unemployment Rate, Percent
0	201507	105.243187	107.083734	103.963874	108.975713	100.477308	606	12.2
1	201508	103.173508	106.724736	98.344205	81.585574	95.429172	619	12.4
2	201509	102.345636	107.566305	100.752634	101.665177	98.535717	633	12.7
3	201510	101.931701	107.662179	101.555444	103.907075	101.253944	639	12.8
4	201511	101.621249	107.450189	101.053688	99.228332	96.885365	634	12.7

Data Profile – Weather Data

Goal: To assess the influence of weather conditions on guests' travel plans and explore any potential correlation with booking cancellations

Characteristics:

1. Average weather condition in Lisbon region between July 1, 2015, and September 30, 2017
2. Includes temperature-related information, humidity values, precipitation, and snow presented as daily averages

	datetime	tempmax	tempmin	temp	dew	humidity	precip	snow
0	2015-07-01	23.8	17.4	19.8	16.5	81.7	2.986	0
1	2015-07-02	26.2	16.0	20.6	13.8	67.0	0.000	0
2	2015-07-03	26.0	16.1	20.9	12.7	61.4	0.000	0
3	2015-07-04	29.2	18.9	22.7	17.7	74.2	0.000	0
4	2015-07-05	27.4	17.8	21.8	12.2	57.2	0.000	0

Data Profile – Tourist Data

Goal: To study monthly travel popularity fluctuations in Lisbon and understand its correlation with our hotel booking data

Characteristics: 'Number of Tourists Data' includes the monthly number of tourist travelling across Portugal border between Jul 2015 and Dec 2017

Number of Tourists Data

	Country	2017M12	2017M11	2017M10	2017M09	2017M08	2017M07	2017M06	2017M05	2017M04	...	2016M04	2016M03	2016M02	2016M01	2015M12	2015M11	2015M10
0	Portugal	142332	135089	151995	212172	345464	269867	193478	132494	215631	...	145539	143662	104442	96123	115301	90633	136812

Data Profile

Main dataset:

Resort & City Hotel Booking

- reservation status, guest, payment, arrival date, etc

Supporting dataset:

Economic Indicator

- CPI, PPI, unemployment rate, etc

Weather Data

- temperature, humidity, snow, etc

Tourism Data

- Tourist(total & stay overnight) number per month

Import the dataset into Jupyter Notebook using Pandas

Explore the DataFrame to understand the data

Get ready for data transformation

```
df_economical = pd.read_excel('economic_data_portugal_1126.xlsx', header=0)
df_economical.head(2)

df_weather = pd.read_csv('portugal_weather_1126.csv', header=0)
df_weather.head(2)

df_tourists = pd.read_excel('tourlists_portugal.xlsx', header=0)
df_tourists
```

jupyter

/ 03

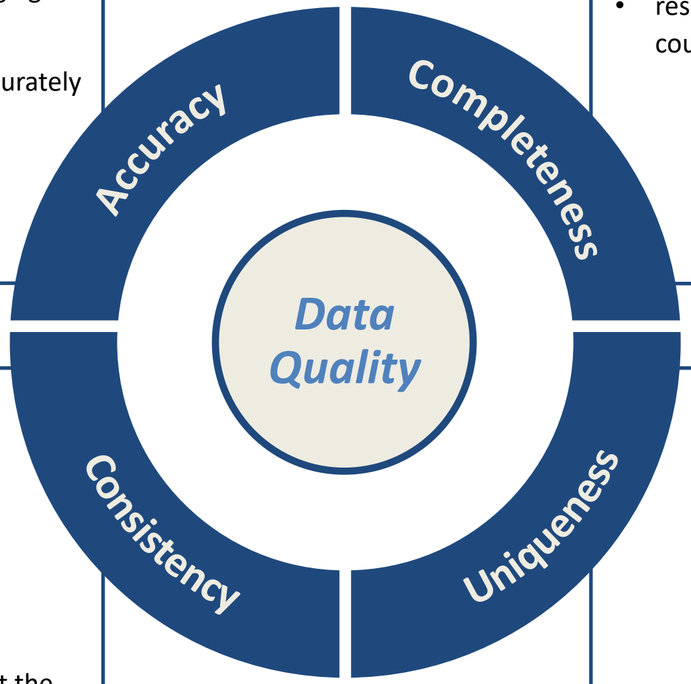
ETL
Data Transformation

Data Assessment

- The data is aggregated simply by summing and averaging the fact data
- Transformation can accurately represent reality

Missing value in

- reservation detail (children, country, agent, and company)



- Data format is NOT consistent through out the dataset (year, month, date)

- NO duplicate data

Data Cleaning – Missing Value

Issue: Missing Value in children(4), country(479), agent(16,126), and company(110,333) columns

Approach: Replace NULL with "0" or "NONE"

```
hotel                               0
is_canceled                         0
lead_time                           0
arrival_date_year                   0
arrival_date_month                  0
arrival_date_week_number            0
arrival_date_day_of_month           0
stays_in_weekend_nights             0
stays_in_week_nights                0
adults                              0
children                            4
babies                              0
meal                                0
country                           479
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                           16126
company                        110333
days_in_waiting_list                0
customer_type                       0
adr                                 0
required_car_parking_spaces         0
total_of_special_requests           0
reservation_status                  0
reservation_status_date             0
arrival_date                        0
date_reservation_status             0
dtype: int64
```

```python
#Data Cleaning
df['children'].fillna(0, inplace=True)
df['country'].fillna('none', inplace=True)
df['agent'].fillna(0, inplace=True)
df['company'].fillna(0, inplace=True)
df = df[df['adr'] > 0 ]
```

Data Cleaning – Combination

Issue: Reservation information spread across 2 hotel tables without a full view.

Approach: Add a "hotel_category" column to identify resort or city type, and combine the 2 tables into one.

```
df_resort= pd.read_csv('df_resort.csv')
df_city= pd.read_csv('df_city.csv')
df= pd.concat([df_resort, df_city], ignore_index=True)
df.head(5)
```

	hotel	is_canceled	lead_time	arrival_date_year	arrival_date_month	arrival_date_week_number	arrival_date_day_of_month	stays_in_weekend_nights	stays_in_week_nights
0	Resort Hotel	0	342	2015	July	27	1	0	0
1	Resort Hotel	0	737	2015	July	27	1	0	0
2	Resort Hotel	0	7	2015	July	27	1	0	1
3	Resort Hotel	0	13	2015	July	27	1	0	1
4	Resort Hotel	0	14	2015	July	27	1	0	2

Data Cleaning – Data Type Standardize

Issue: Time format are different across 4 datasets

Approach: Change to YYYY-MM-DD format

	arrival_date_day_of_month	reservation_status_date
0	1	2015-07-01
1	1	2015-07-01
2	1	2015-07-02
3	1	2015-07-02
4	1	2015-07-03
...	...	...
119385	30	2017-09-06
119386	31	2017-09-07
119387	31	2017-09-07
119388	31	2017-09-07
119389	29	2017-09-07

```
#Hotel Arrival Data
df['arrival_date_year'] = df['arrival_date_year'].astype(str)
df['arrival_date_month'] = df['arrival_date_month'].astype(str)
df['arrival_date_day_of_month'] = df['arrival_date_day_of_month'].astype(str)

df['arrival_date'] = pd.to_datetime(df['arrival_date'])

df['arrival_date'] = df['arrival_date_year'] + '-' + df['arrival_date_month'] + '-' + df['arrival_date_day_of_month']
df
```

```
#Hotel Reservation Data
#Hotel Reservation Status Data
#Tourist Data
```

/ 04

ETL
Data Loading

Data Loading

Export the cleaned data to CSV file

```python
# hotel_tourism_link
selected_columns = [
    'booking_id', 'tourism_id'
]

# Creating a new DataFrame with only the selected columns
selected_df = df_combined[selected_columns]
# Saving the DataFrame to a CSV file
selected_df.to_csv('hotel_tourism_link.csv', index=False)


# hotel_weather_link
selected_columns = [
    'booking_id', 'weather_id'
]
# Creating a new DataFrame with only the selected columns
selected_df = df_combined[selected_columns]
# Saving the DataFrame to a CSV file
selected_df.to_csv('hotel_weather_link.csv', index=False)
```

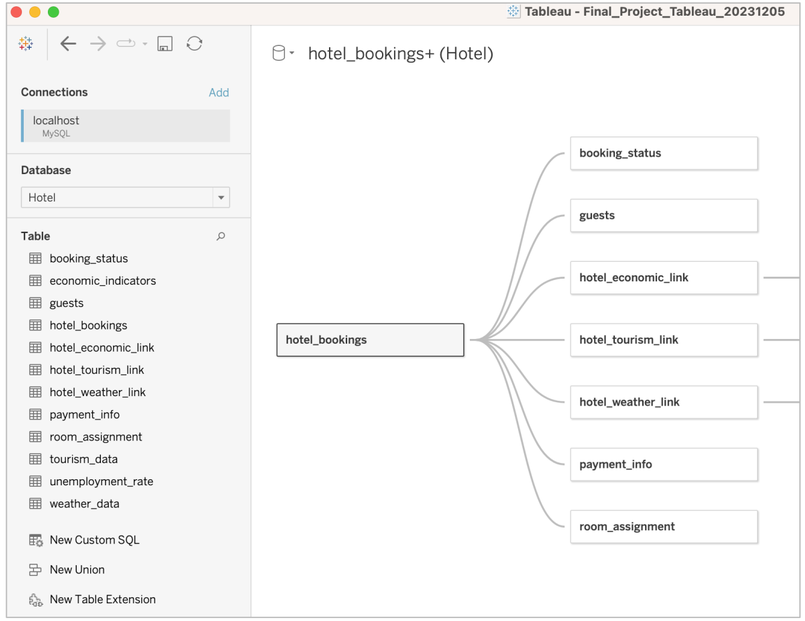
Load data into MySQL Workbench

```sql
CREATE DATABASE IF NOT EXISTS Hotel;
USE Hotel;

CREATE TABLE hotel_bookings (
    booking_id INT PRIMARY KEY,
    hotel VARCHAR(20),
    lead_time INT,
    arrival_date DATE,
    stays_in_weekend_nights INT,
    stays_in_week_nights INT,
    meal VARCHAR(20),
    country VARCHAR(20) NULL,
    agent VARCHAR(20) NULL,
    company VARCHAR(20) NULL,
```

```sql
LOAD DATA LOCAL INFILE '/Users/sss/Desktop/DE_Hotel Project/hotel_bookings.csv' IGNORE
INTO TABLE hotel.hotel_bookings
FIELDS TERMINATED BY ','
OPTIONALLY ENCLOSED BY '"'
lines terminated by '\n' STARTING BY ''
IGNORE 1 ROWS;
SELECT * FROM hotel_bookings;
SELECT COUNT(*) AS total_rows FROM hotel_bookings;
```

Connect to Tableau



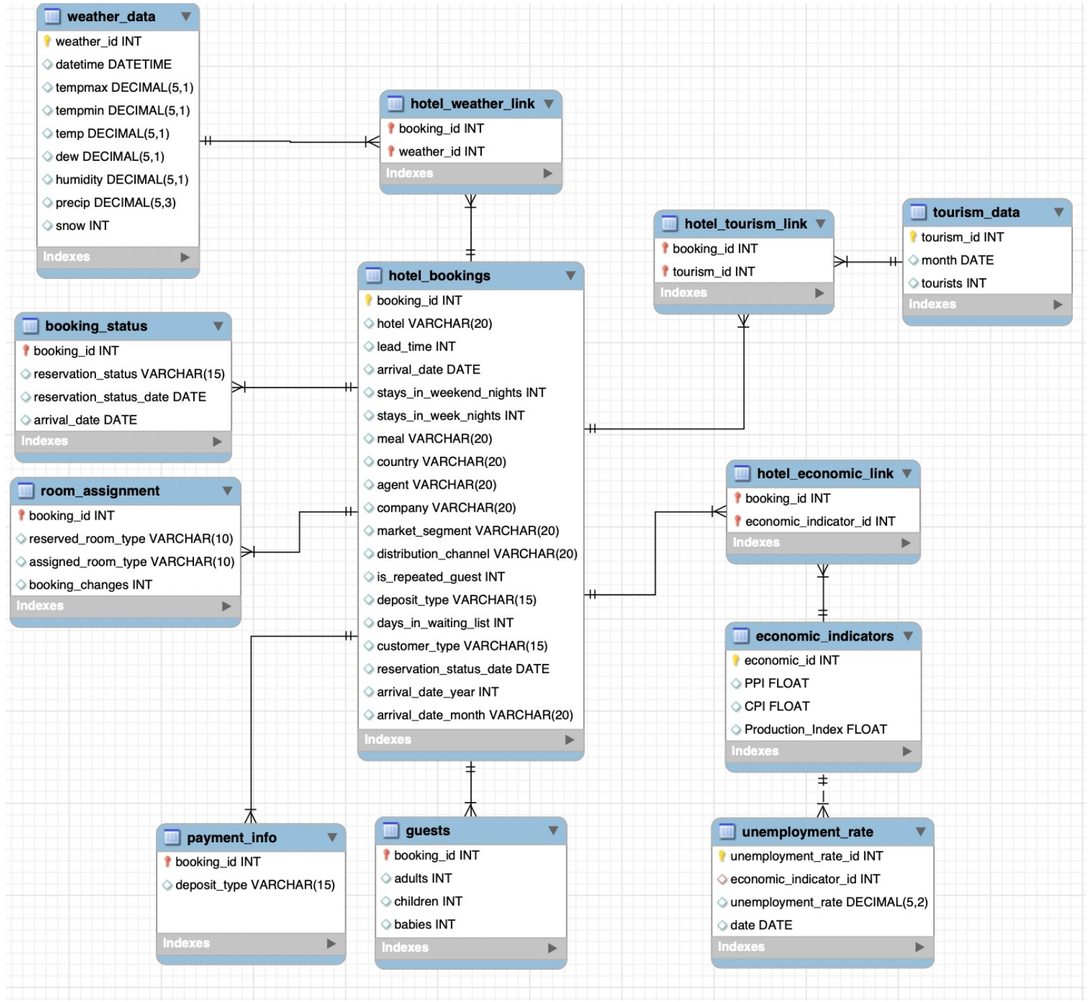
Data Modeling

1. After collecting the data from various sources, it was **converted into tables** to adhere to the **Entity-Relationship Model** (OLAP).

2. Since the most important parameter was the **date_id**, we included booking, arrival, and status change dates in a table called **Hotel_Booking**. This table is also the core of our snowflake, contains information for all key dates, booking details, and statuses governed by the booking_id.

3. Our focus was on **4 sectors**: Booking Detail, Economics, Weather, and Tourism. Therefore, we built tables for each of these sectors, normalizing them to contain atomic information without any functional and transitive dependencies.

EER



- Start with the EER model in MySQL
- Each table contains a unique key as its primary key.
- Foreign keys, such as booking_id, tourism_id, weather_id, and economic_indicator_id, are used to establish relationships between the tables.
- There are several lookup tables: Room_Assignment, Payment_Info, Booking_Status, Guests, Tourism_Data, Economics_Indicators, Unemployment_Rate, and Weather_Data.

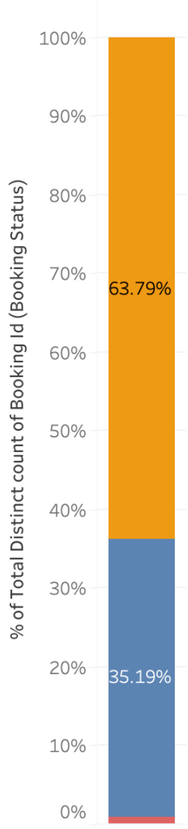
/ 05

EDA and Visualization

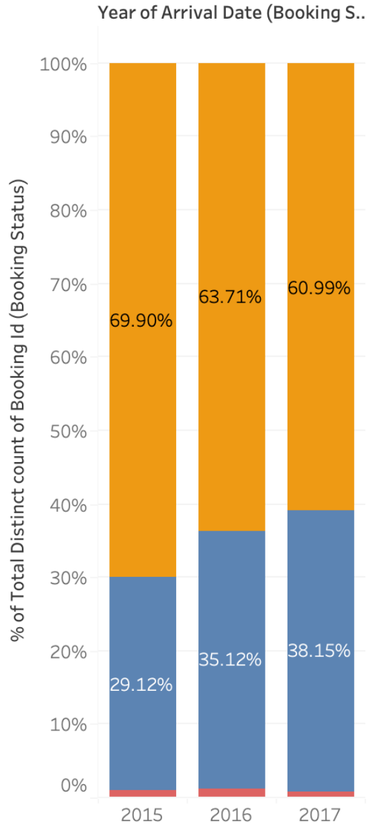
Cancellation Trend

- The average cancellation rate is 35.19%.
- There is a growing trend of cancellation rates, increasing from 29% in 2015 to 38% in 2017.
- There is seasonal fluctuation, with a higher cancellation rate from April to June; The seasonal fluctuation remained consistent across the years.

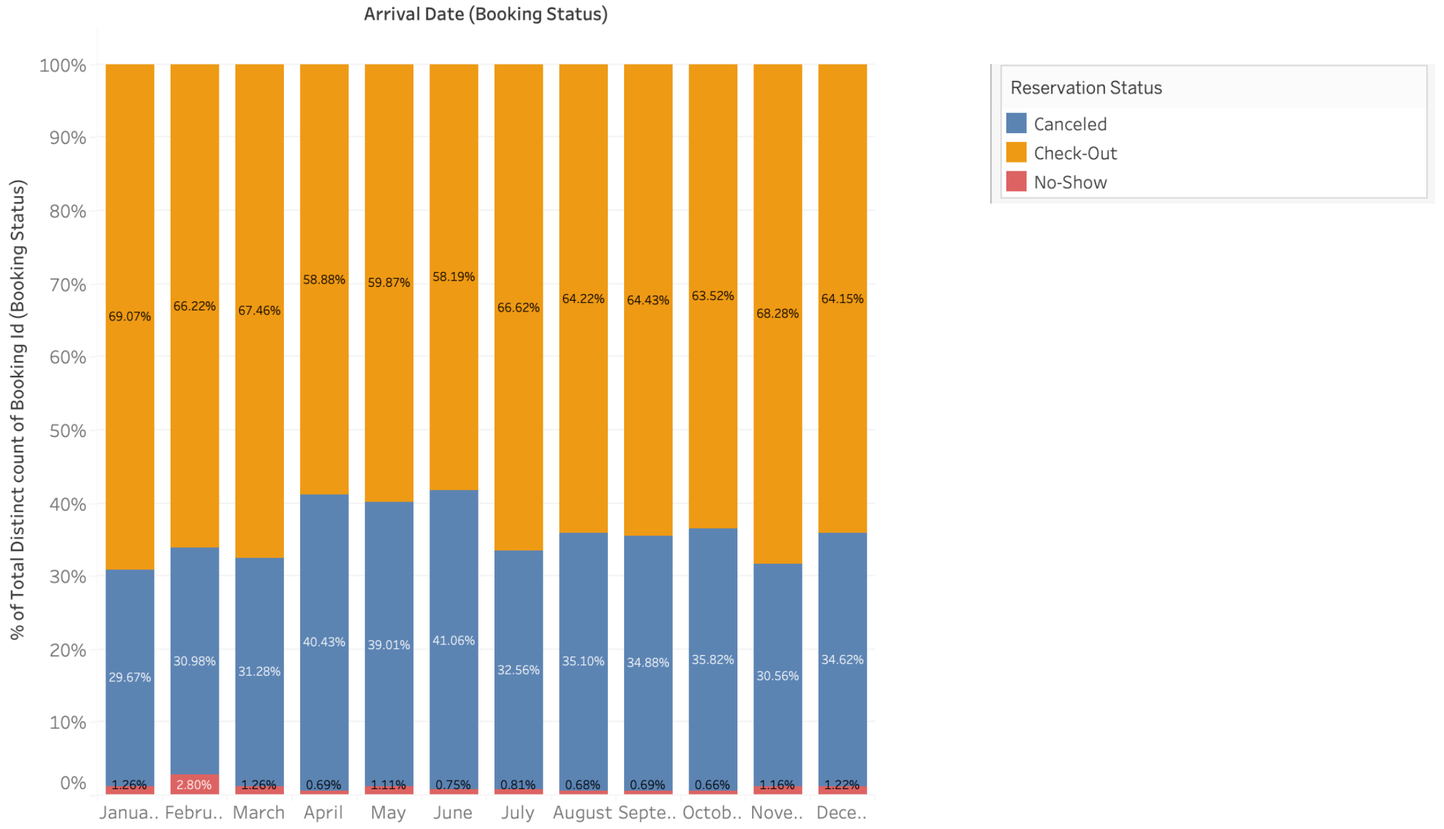
Cancel Rate



Cancel Rate Trend by Year

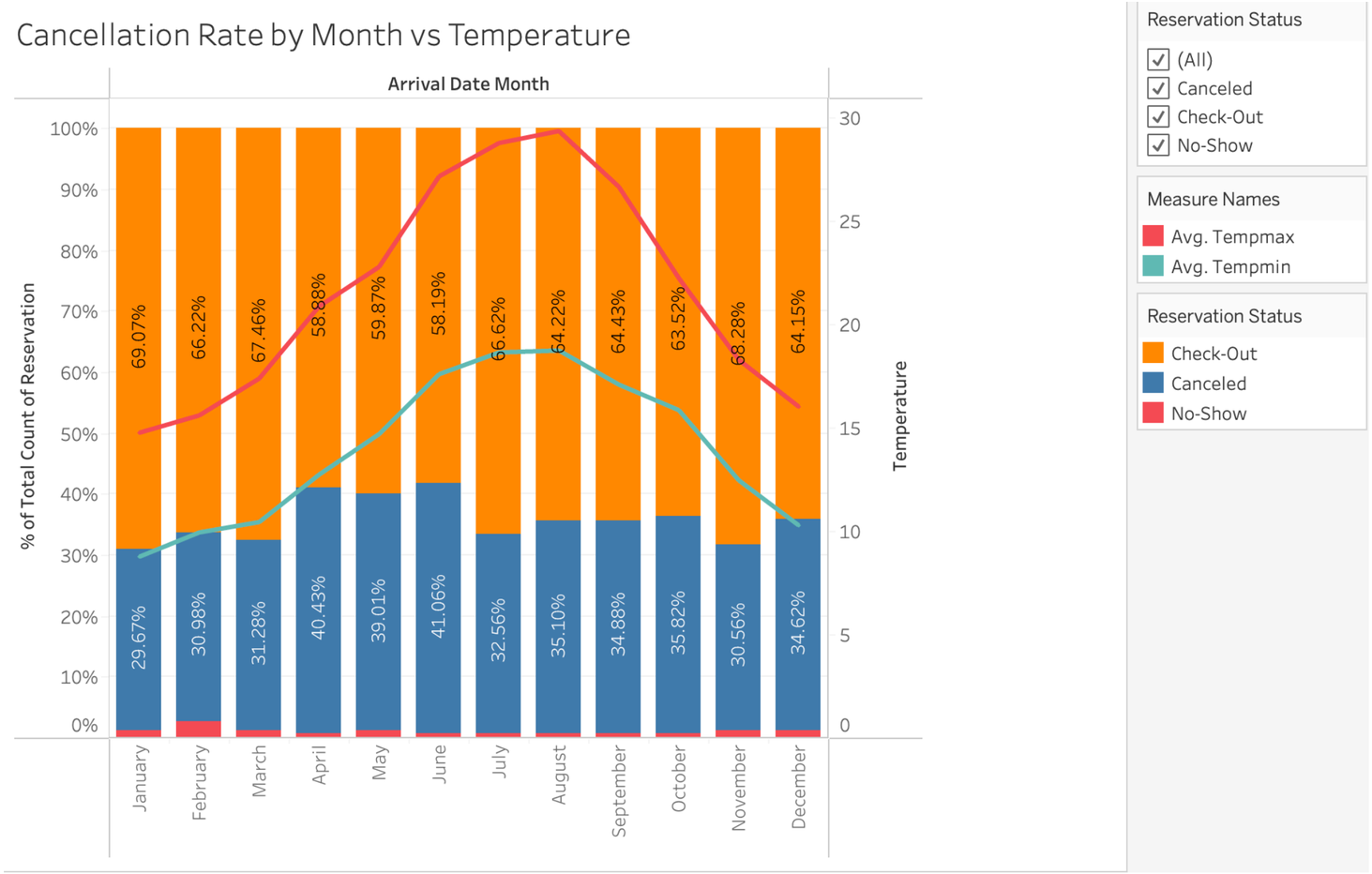


Cancel Rate Trend by Month



Cancellation Fluctuation by Temperature

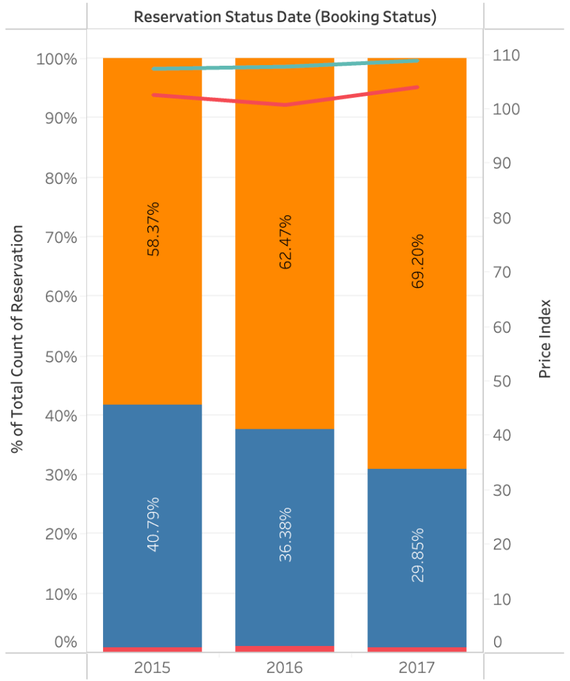
- Temperature does not have a significant relationship with the cancellation rate.



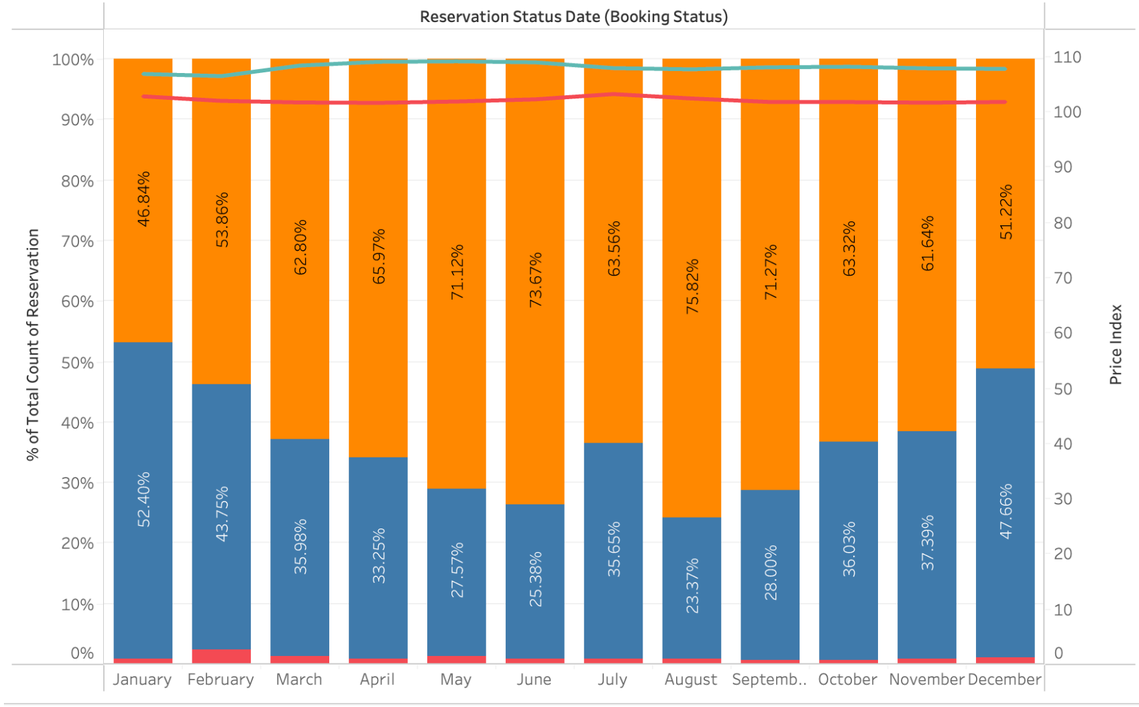
Cancellation Fluctuation by Economic Environment

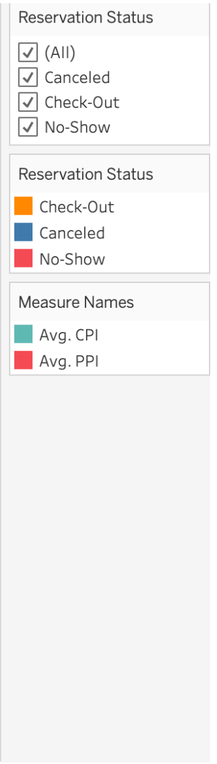
- CPI and PPI do not show significant correlation with the cancellation rate in the year dimension.
- In the month dimension, there is a negative correlation between PPI and cancellation rate.

Cancellation Rate by Year & Economical Condition



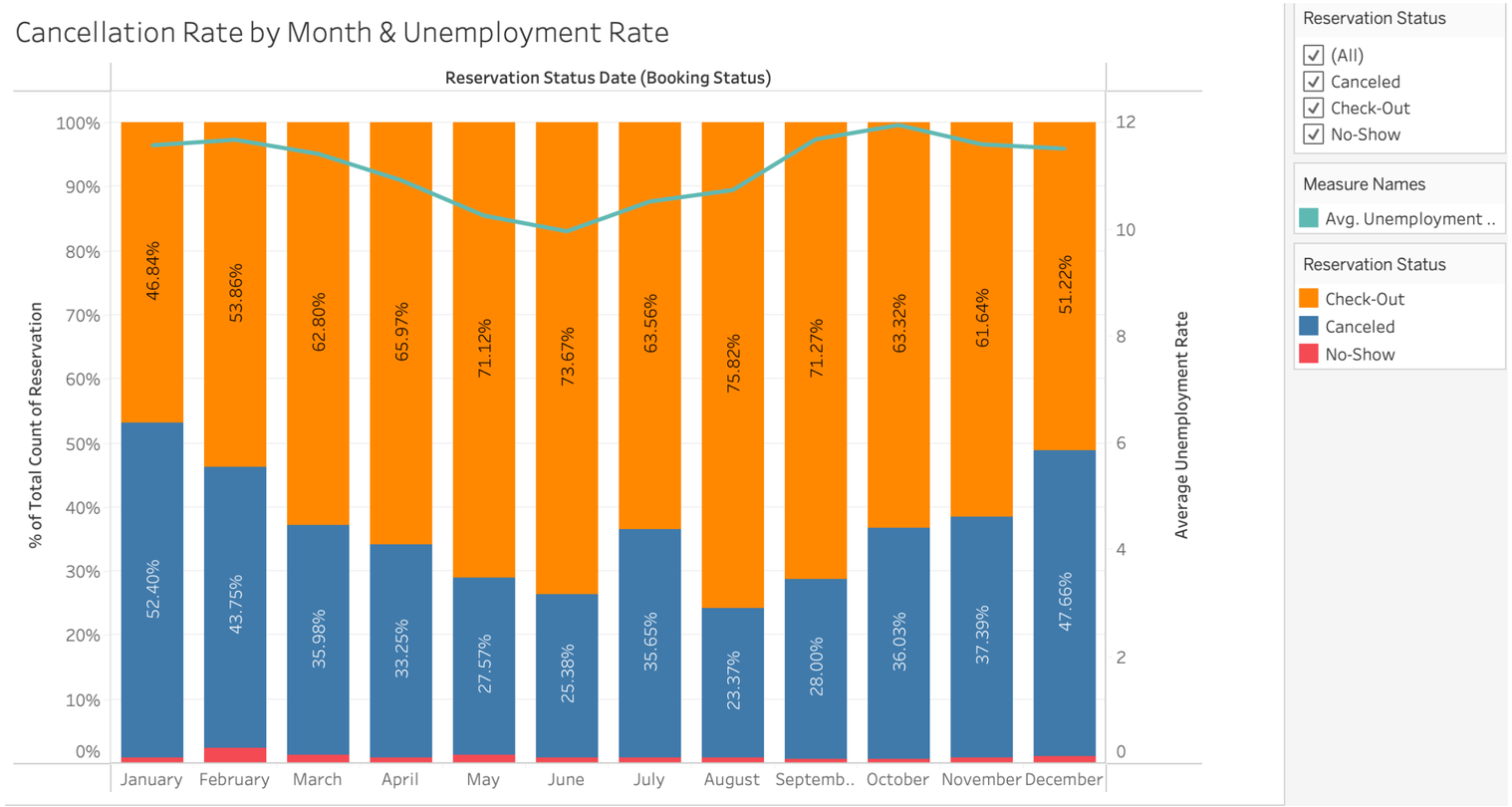
Cancellation Rate by Month & Economical Condition





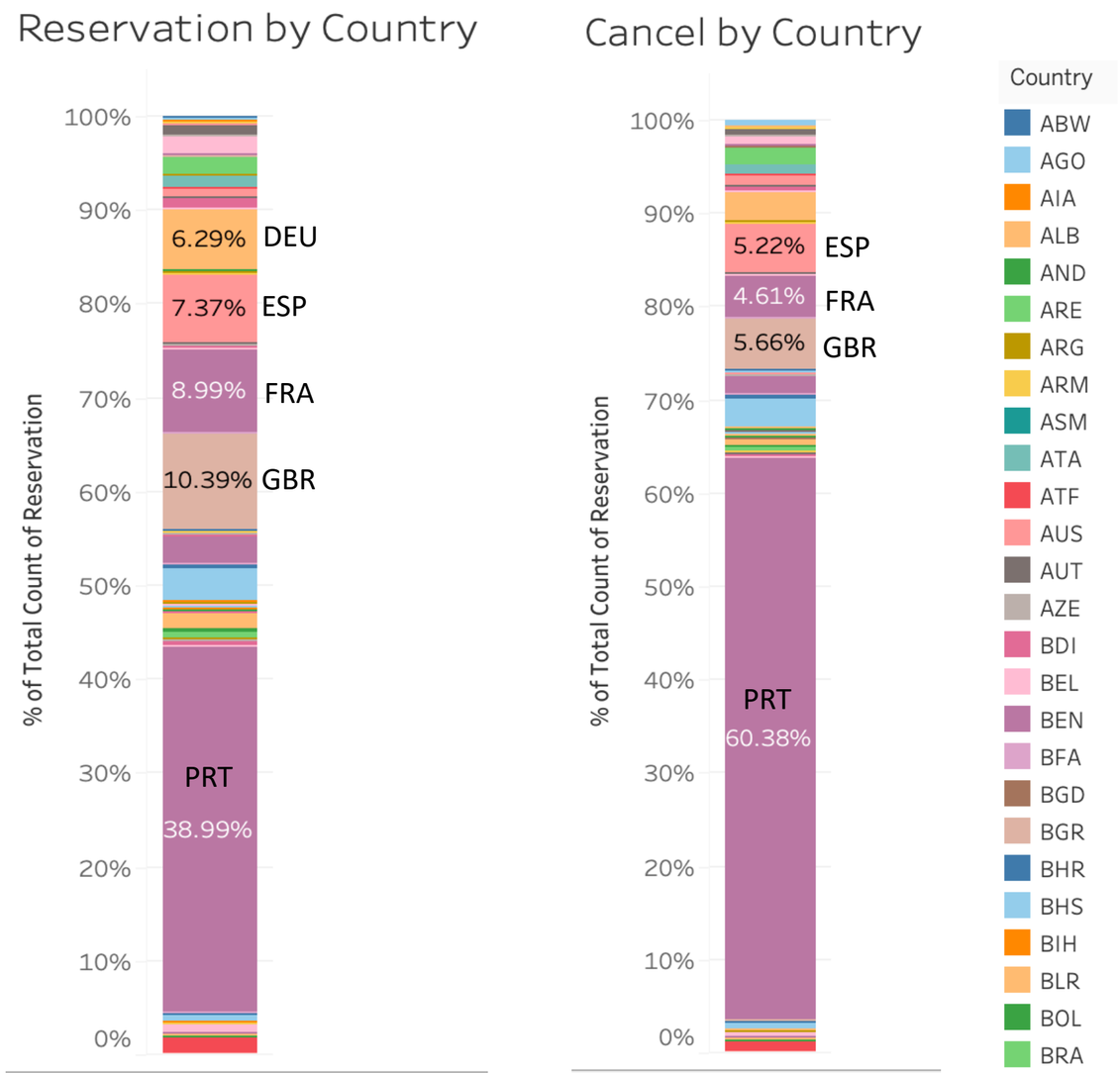
Cancellation Fluctuation by Economic Environment

- The unemployment rate has a positive relationship with cancellation rate.



Country of Origin

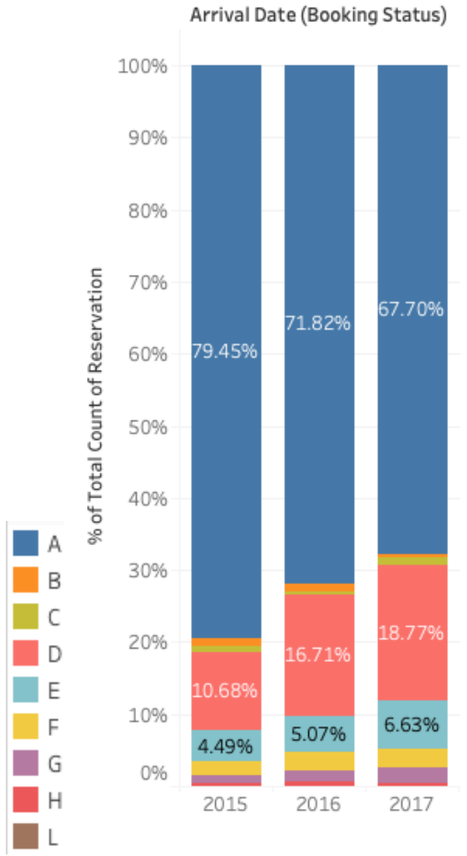
- The cancellation rate of PRT (Portugal) is significantly higher than that of other countries. (Domestic traveler might have a higher flexibility on travel plan)



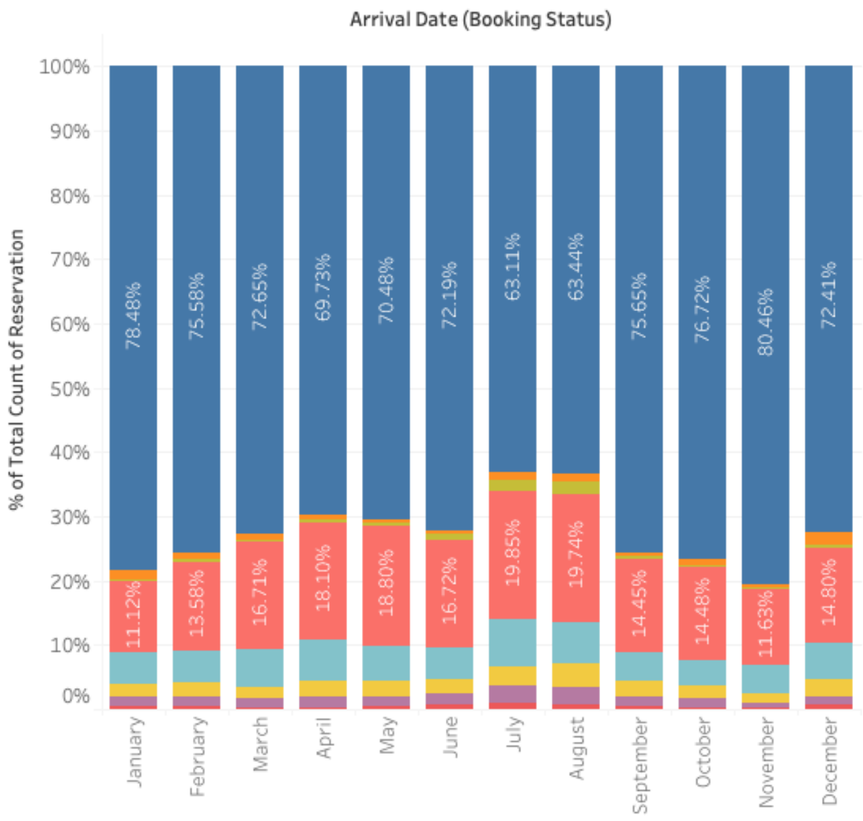
Room Type

- Room types A, D, and E are the most popular, with not much difference in seasonality.
- The reservation proportion of room type A has decreased, while the proportions of room types D, E, F, and G have increased.
- The cancellation rate for each room type ranges from 27% to 39%. Room types H and A have the highest cancellation rates.

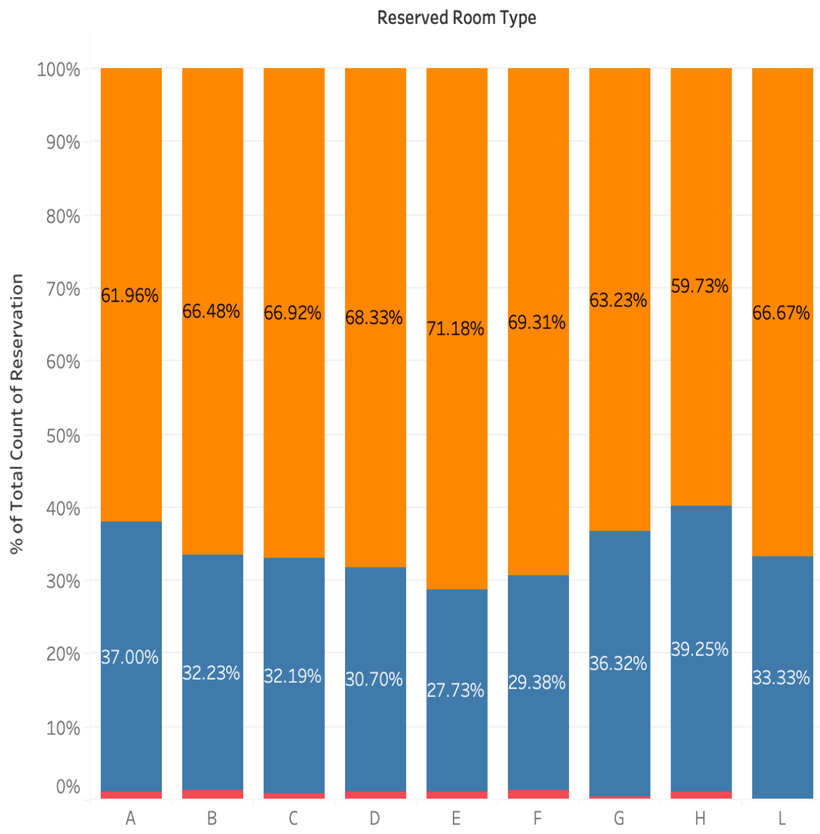
Yearly Reservation% by Room Type



Monthly Reservation% by Room Type



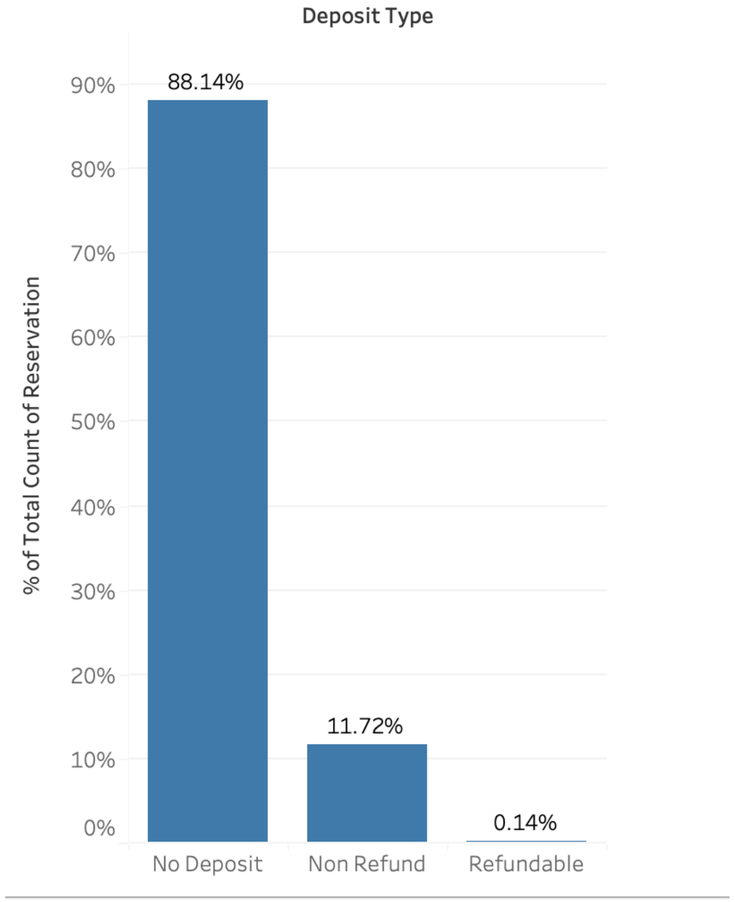
Cancellation Rate by Room Type



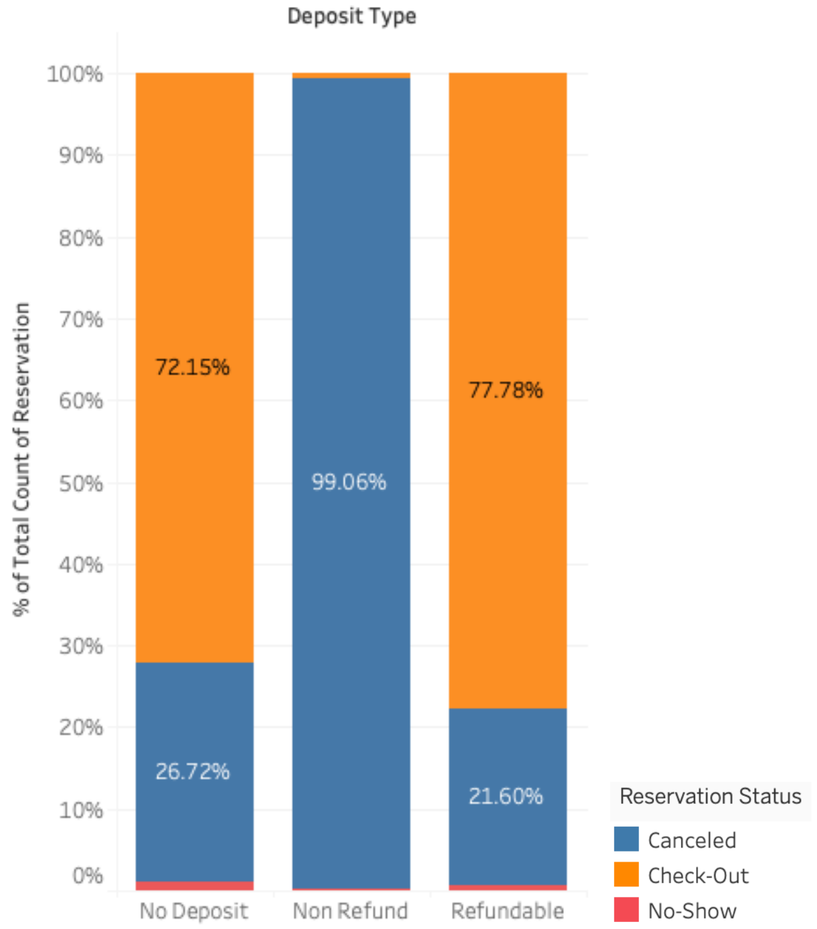
Deposit Type

- 88% of reservations do not require a deposit, while 12% is non-refundable
- The non-refundable payment reservations have the highest cancellation rate at 99% (Usually, the cheapest room will have a no-refund policy)

Number of Reservation by Deposit Type



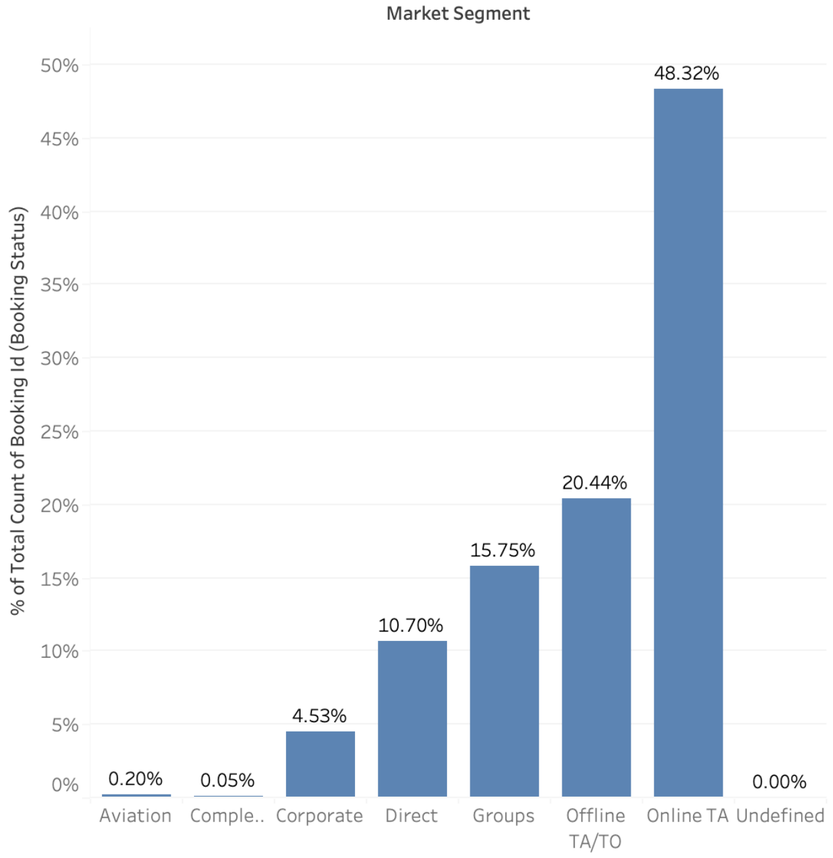
Cancellation Rate by Deposit Type



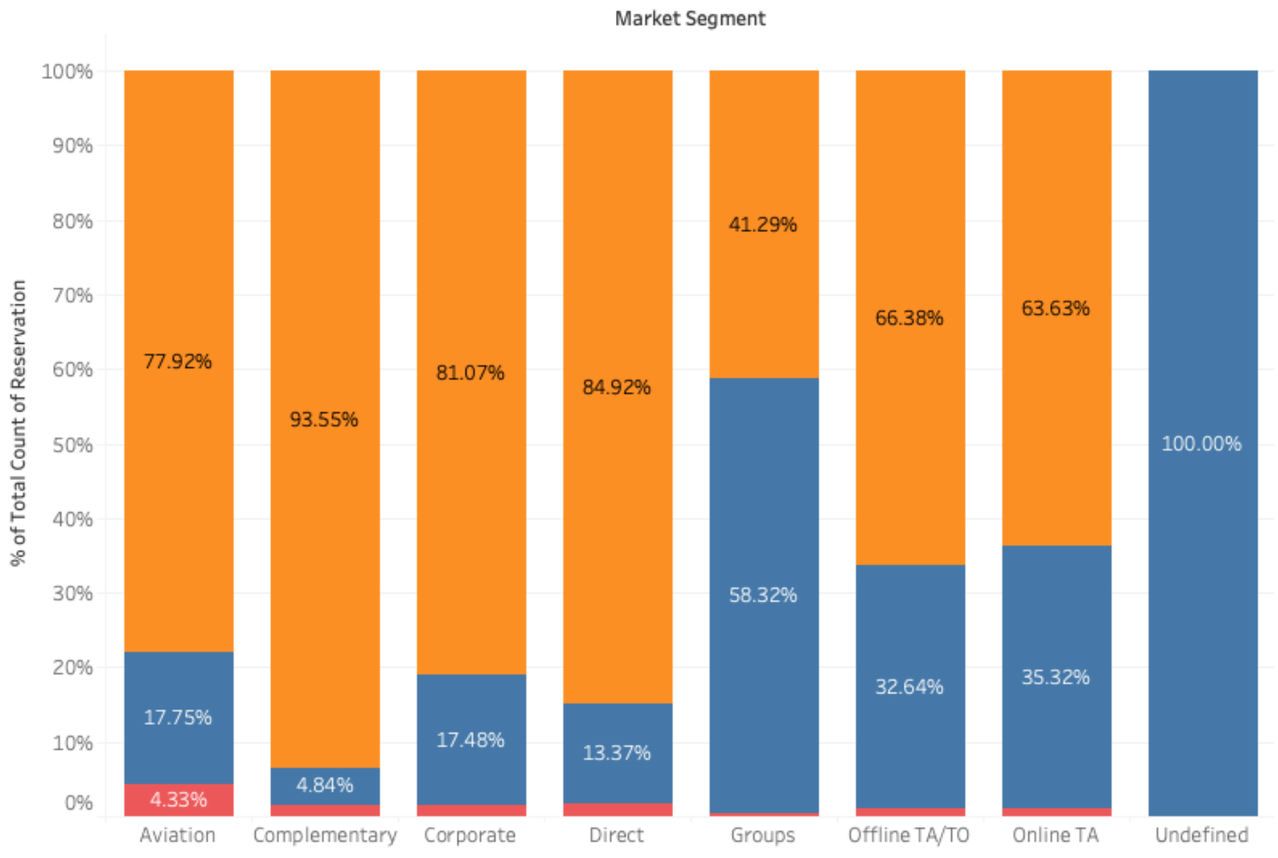
Market Segment

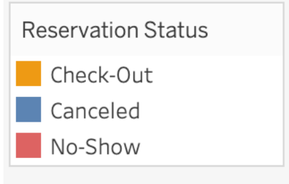
- Almost half of the customers are located under the Online Travel Agent segment.
- The Group segment has the highest cancellation rate at 58%, for TA/TO is about 40% (Hotels usually provide a flexible cancellation policy for group & agent clients, while group & agent also tend to make bookings in advance to secure a certain number of rooms at negotiated rates)

Reservation by Mkt Segment



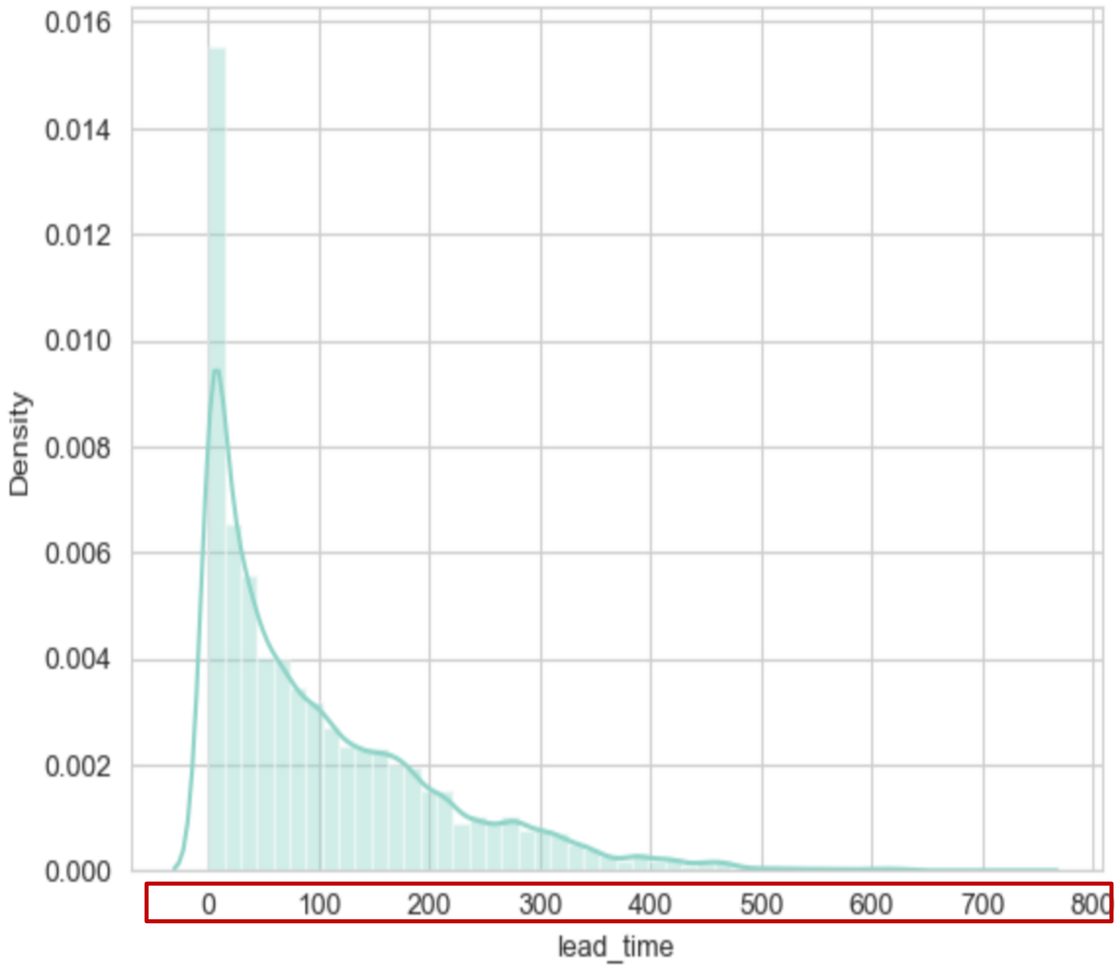
Cancellation% by Mkt Segment





Reservation Leadtime

- The lead time between the room booking date and the arrival date are right skewed, vary from 0 to 800 days.
- The longer the lead time is, the higher the potential for the order to be canceled.



Cancellation Orders by Leadtime

gp	count	sum	ratio
3	20647	119390	17.29%
2	26713	119390	22.37%
1	72030	119390	60.33%

Cancellation Rate by Leadtime

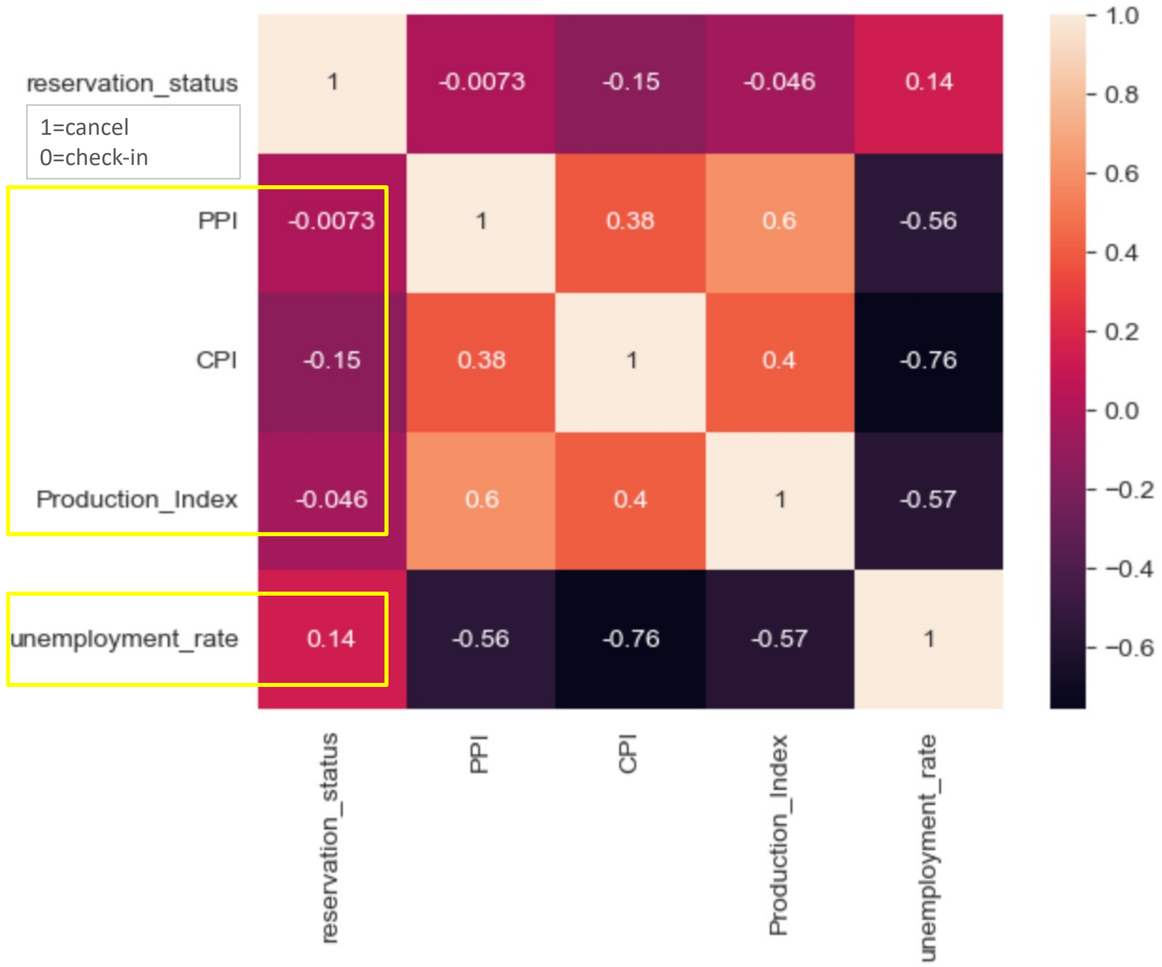
gp	1	0
3	60.33	39.67
2	22.37	77.63
1	17.29	82.71

1=cancel
0=check-in

Gp1= leadtime < 100 days
Gp2= 100< leadtime < 200 days
Gp3= leadtime > 100 days

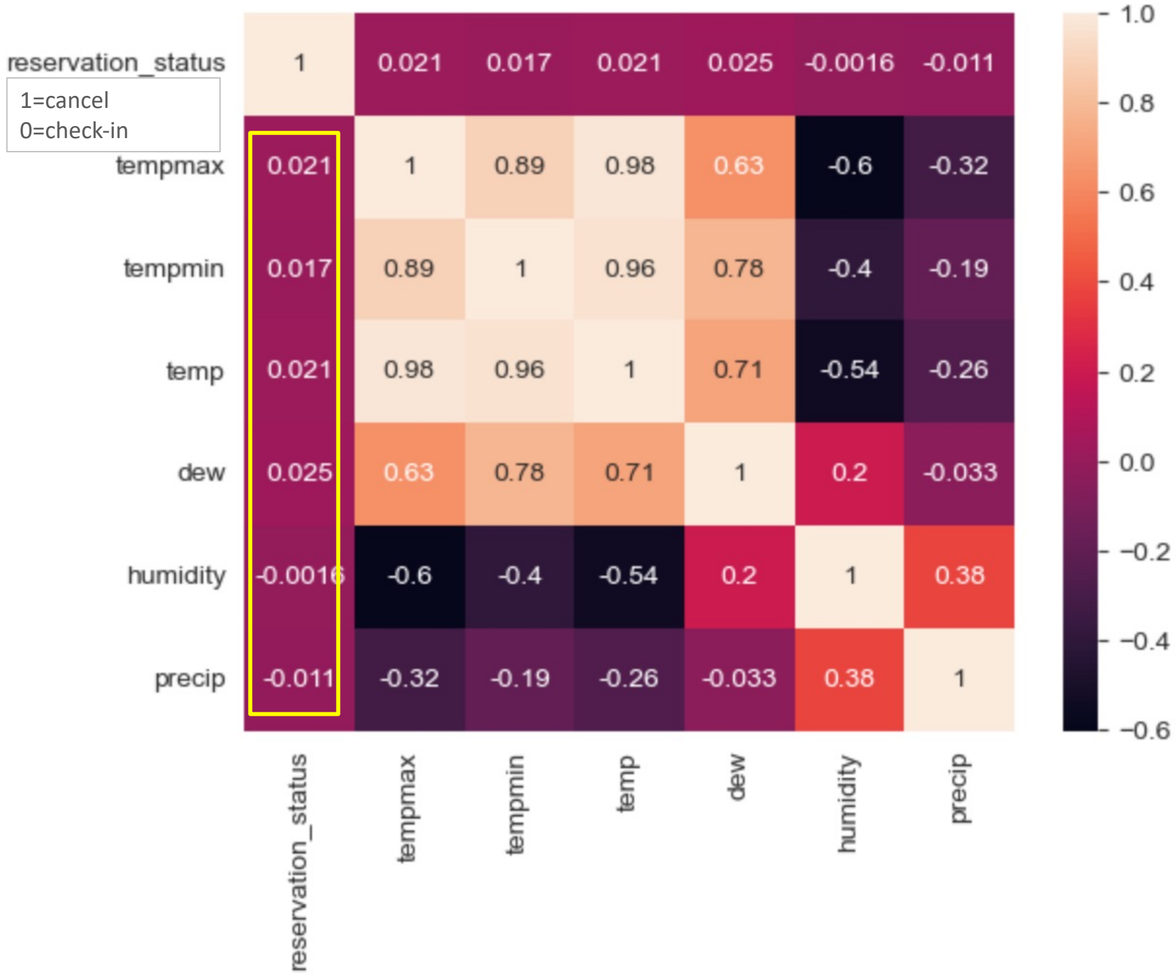
Correlation between Cancellation Rates & Economic Environment

- Cancellation rate has a negative correlation with economic factors such as CPI, Production Index, and PPI.
- Cancellation rate has a positive correlation with unemployment.



Correlation between Cancellation Rates & Weather

- The relationship between weather and cancellation rate is insignificant.



/ 06

Machine Learning

Machine Learning: Random Forest

[Preparation for Modeling]

- Train size = 0.7
- Test size = 0.3
- Encoder: categorical list was label encoded(to digits)

Random Forest

Max depth: 6

Number of trees:400

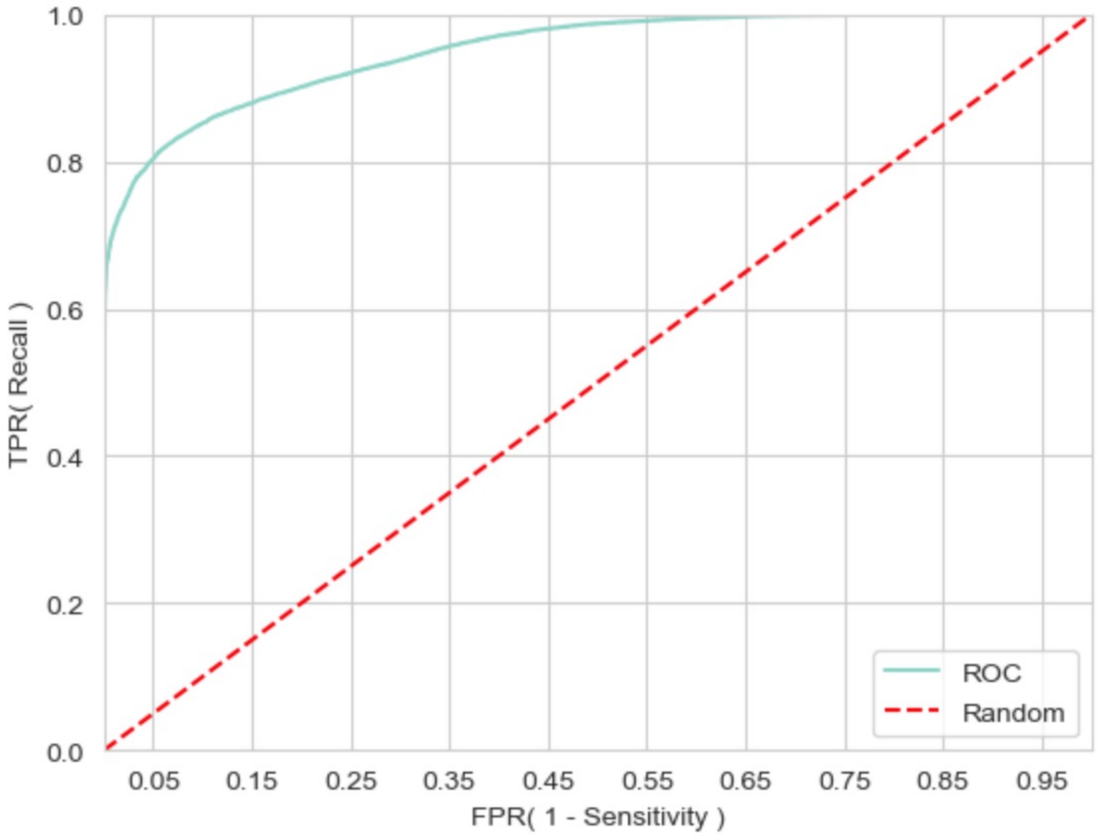
Balanced Training

Precision
84%

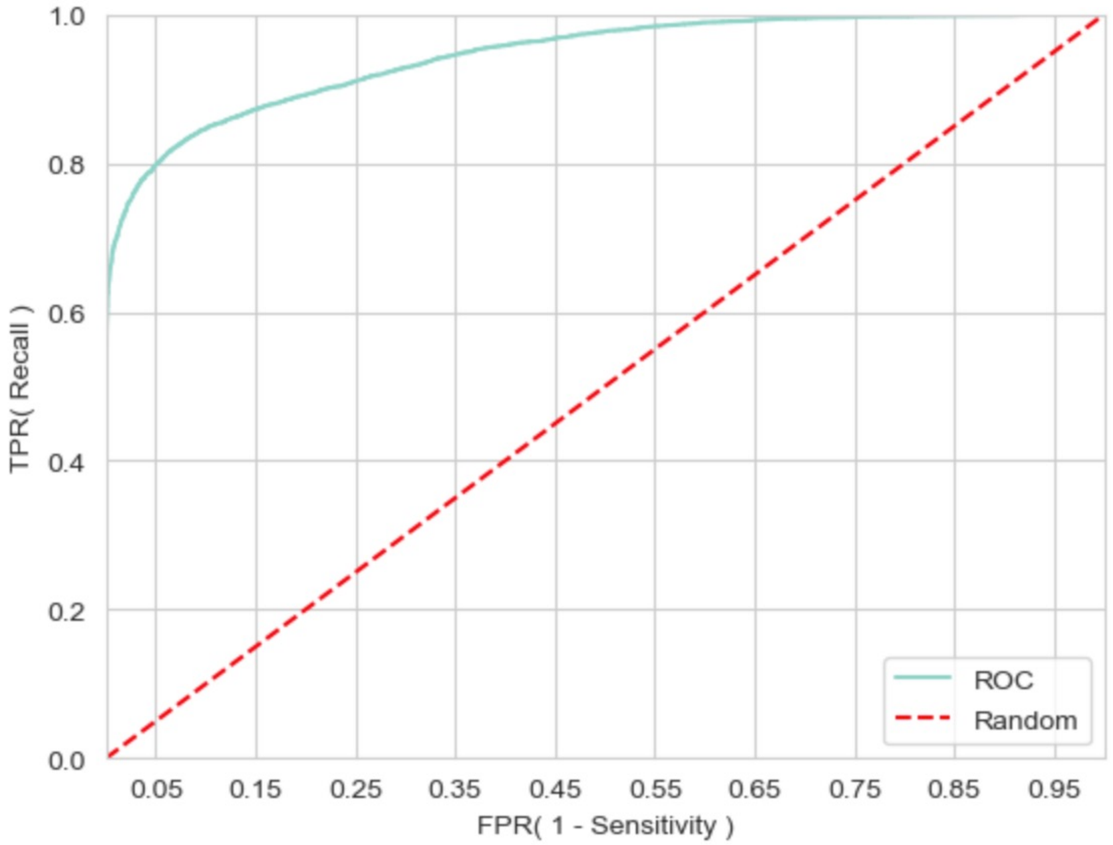
From the ML, got good result of Precision, But this model should be assessed by other way further

Machine Learning: Assessment ROC Curve

```
roc_curve_plot(y_train, y_pred_train_proba)
```



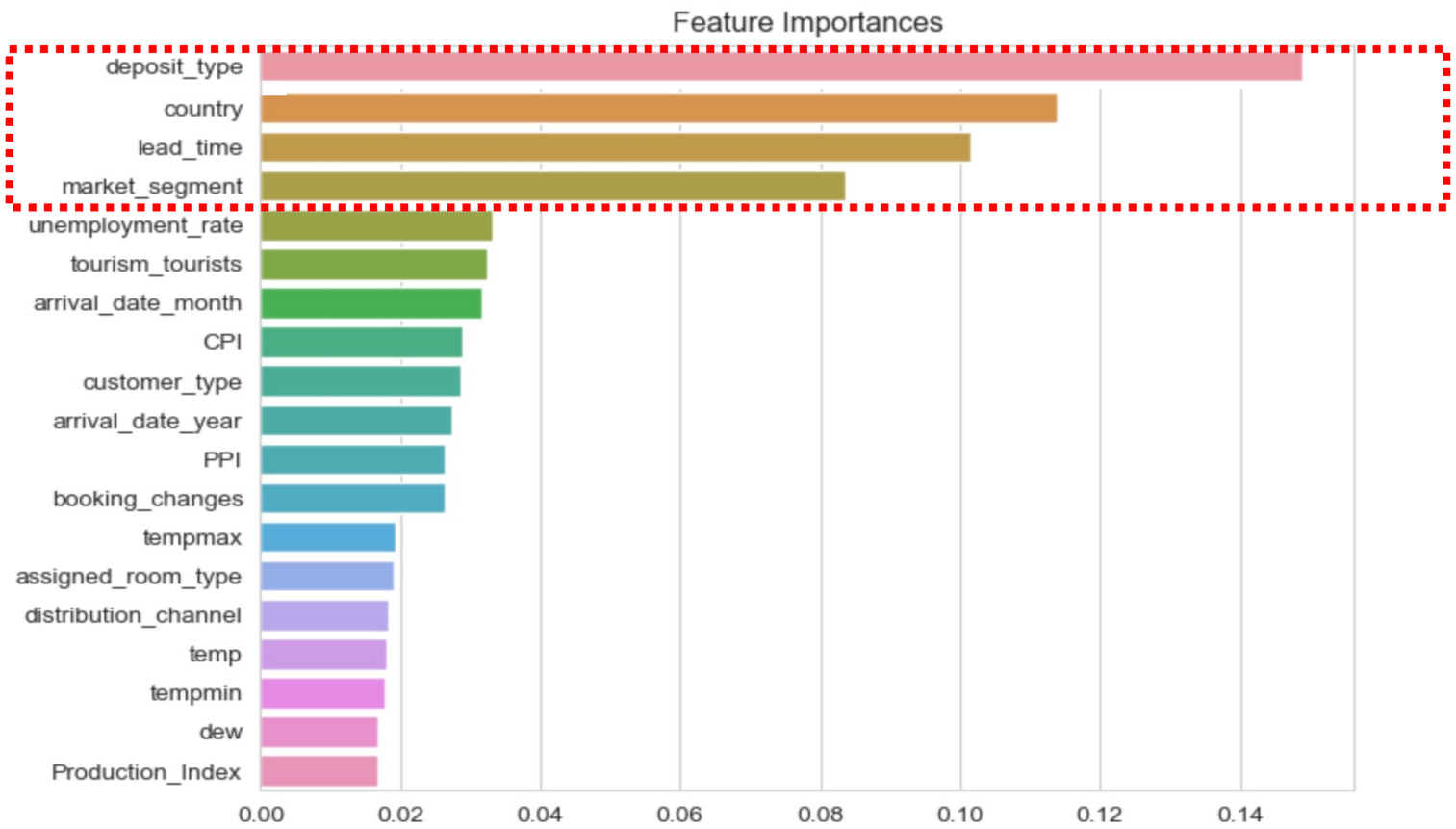

```
roc_curve_plot(y_test, y_pred_test_proba)
```



This model AUC 92% score of ROC curve for both. We expedite this model for further Analysis

Machine Learning: Feature Importance

- The significant features that impact the cancellation rate are deposit type, country, lead time, and market segment.
- External factors have a weak impact on the cancellation rate.

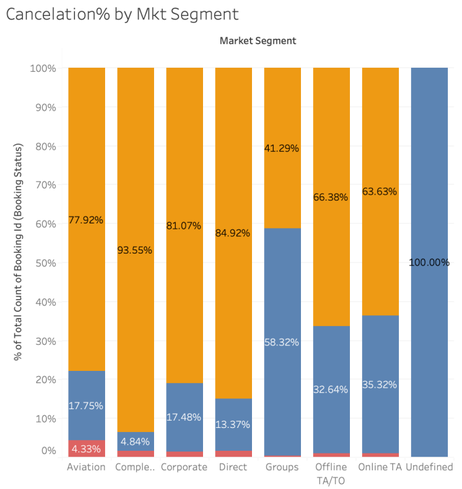


From the ML, these 4 features are most effective for Cancelation rate. Why is it related?

/ 07

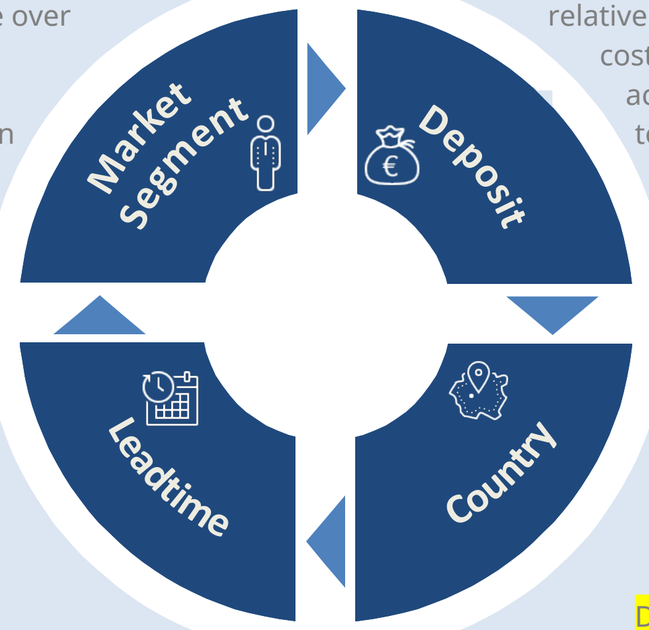
Recommendations and Lessons Learn

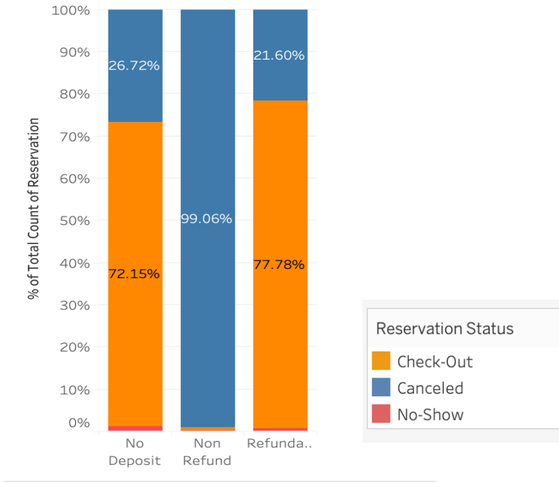
Insights



Groups and TA/TO have higher cancellation rates as these agents buy bulk tickets upfront, meanwhile hotel provides more flexible cancellation policies in contact. The over pre-order and low opportunity cost lead to a higher cancellation rate.

No-Deposit & Refundable has a lower cancellation rate than non-refundable. Most of the time, the cheaper offer will have a no-refund policy. As the price is relatively low, the opportunity cost for clients to cancel is acceptable, which leads to a higher cancellation rate.

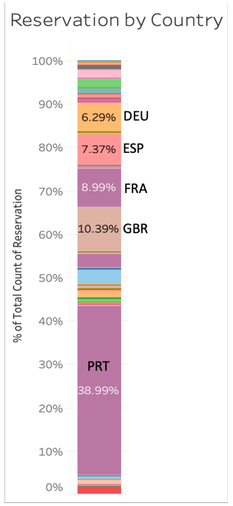


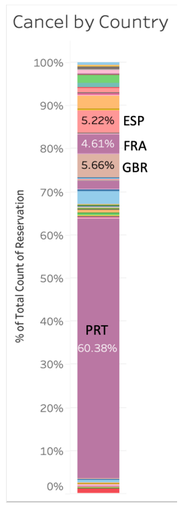


gp	1	0
3	17.29	39.67
2	22.37	77.63
1	60.33	82.71

The length of lead time is positively related to the cancellation rate as it is difficult to predict future travel.

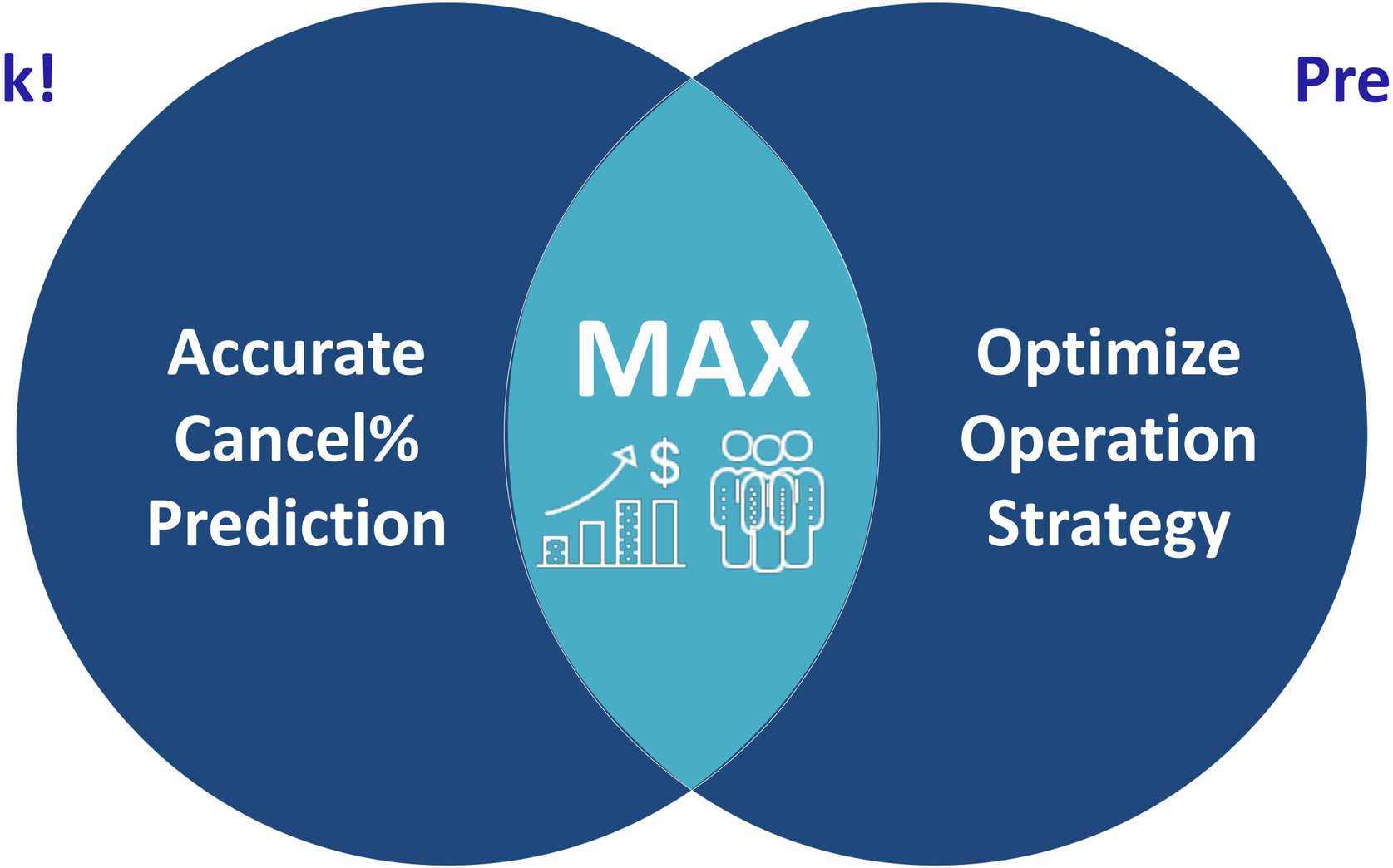
Domestic clients have a higher cancellation rate as they are more flexible with travel plans and the travel costs are relatively low compared to international clients.





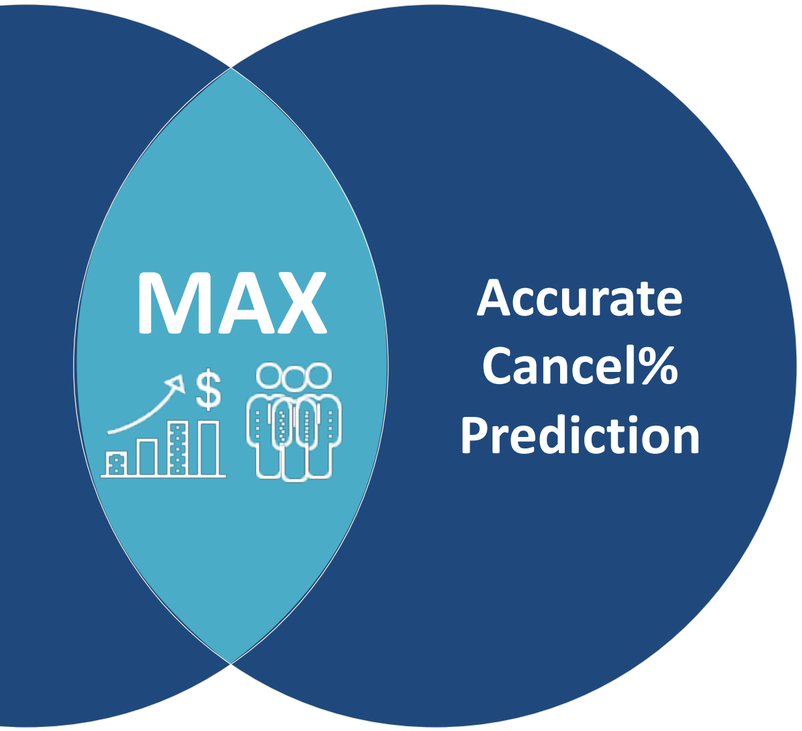
Strategy

Overbook!



Precision TA!

Strategy



CANCEL

consider as Ghost Client
Increase **OVERBOOK** rate

CHECK IN

Precision
84%

Strategy

Deposit

- Mission: Increase **opportunity cost**
- Approach: Change pricing to mid-price with a deposit (consumers will only make orders with serious consideration)
- Jobs to be done: (1) Dive into cancellation and price correlation tests (2) Test consumer price elasticity

Leadtime

- Mission: Maximize hotel **occupancy** rate
- Approach: Consider long lead time cancellations and near-to-date order demand and calculate the optimal cancellation due day to maximize occupancy
- Job to be done: (1) Research on realization approach

Country

- Mission: Increase the **visibility** among **international** travelers
- Approach: Secure higher placement through paid bids on Google Search and TA/TO search results for searches originating from websites outside Portugal
- Jobs to be done: (1) SEM optimization (2) Implement bid management strategies for paid advertising

Segment

- Mission: Increase **reliable client** group’s proportion
- Approach: Cooperate with Corporate, Aviation, and Complimentary clients and review contracts signed with Group and TA/TO
- Jobs to be done: (1) Strengthen cooperation BD (2) Review agency contract details

Thank You

Kitae Kim, Reina Chen, Sally Lee