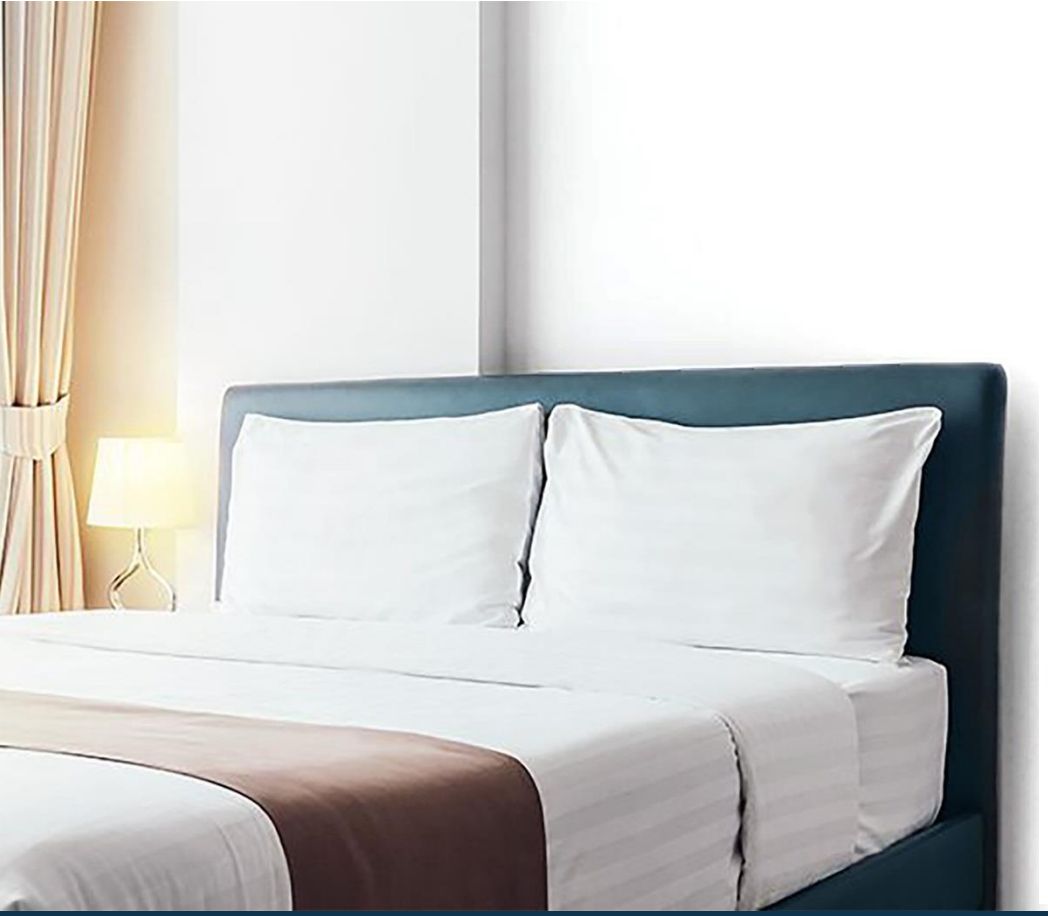


Hotels Booking Cancelation Prediction

Yang Fei

Mentor: Kenneth Gil-Pasquel

Data Science Capstone Project 2, July 2020

What is the Target?

Cancellation Prediction

Hotels

potential trends
and interesting relationships

Customer

Booking Information

Who might cares?

Hotel Management

With prediction, hotel management could propose some overall arrangements and backup measures to reduce the influence caused by the financial loss.

Tourists

By observing the trend of cancelation of a hotel, tourists could know the probability of booking a hotel on their desired dates.

Hotel Industry Researchers

By researching the cancellation prediction, researchers could summarize the developing trends and regulations of this industry

Online APP Developer

Combine the prediction into their booking product to make it more competitive

Where is the data from?

Raw Data

119390 rows and 32 columns

The data is originally from the article [Hotel Booking Demand Datasets]

17 Numerical Variables
17 Categorical Variables
2 Dummy variables

- Deal with missing data
- Replace and drop the ‘undefined’
- Detecting & Filtering Outliers

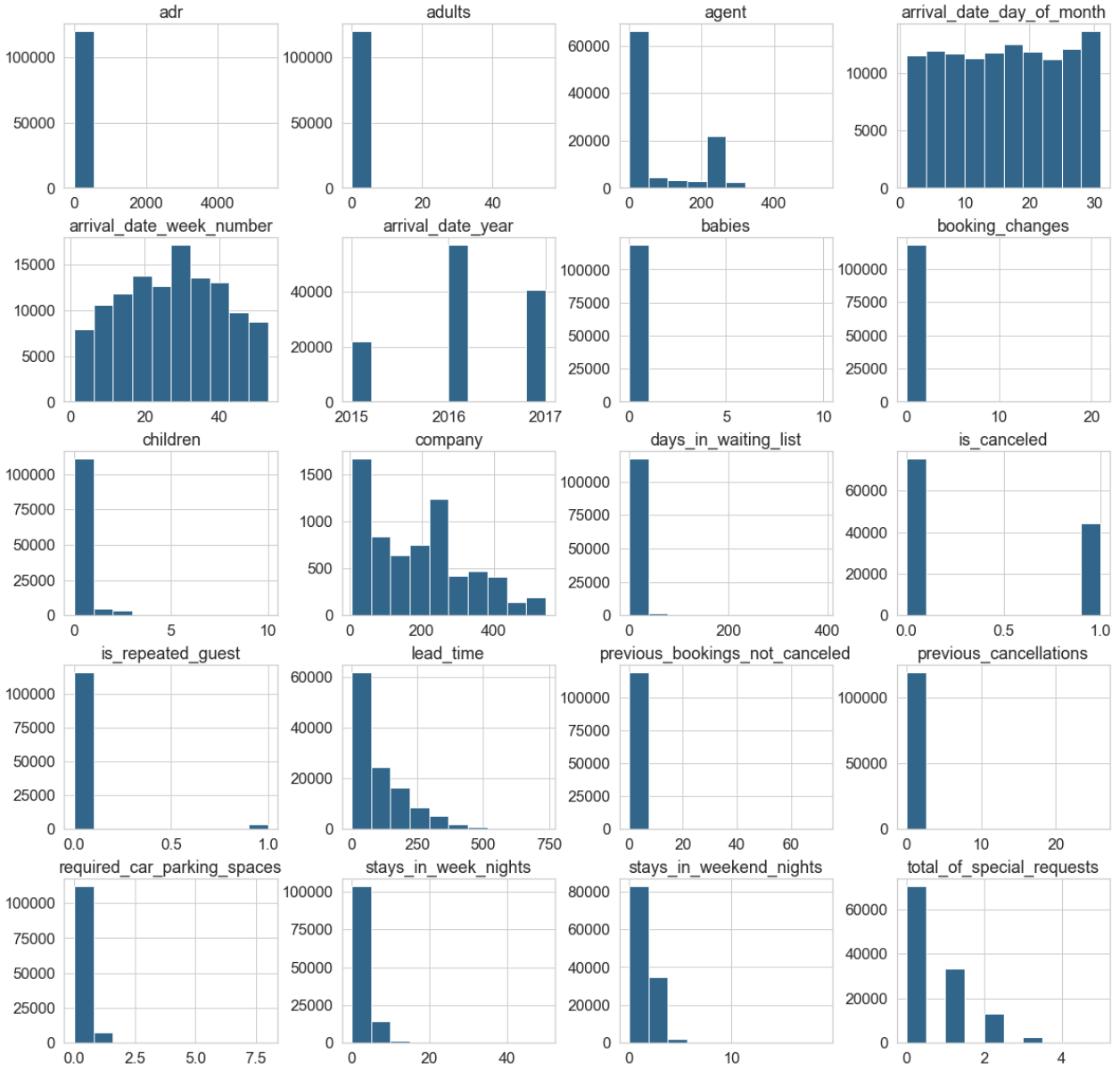
DATA CLEANING

Clean Data

115595 rows and 25 columns

Some columns had a large proportion of missing data.
So dropped them with the unnecessary columns together.

In analysis part I will add 3 features:’week/weekend’, ‘total_span_of_stay’, ‘number of kids’



Data Analysis Mind Map

EDA

- General Analysis
- Target Related Analysis

Statistical Analysis

mathematic relation related to target

Features	Research Objects
All Features	Booking Number Cancellation Number Cancellation Rate
Numerical Features	Cancellation Rate

VS

$$\text{Cancellation Rate} = \frac{\text{Cancellation Number}}{\text{Booking Number}}$$

Important Relationships

EDA- When is the peak season?

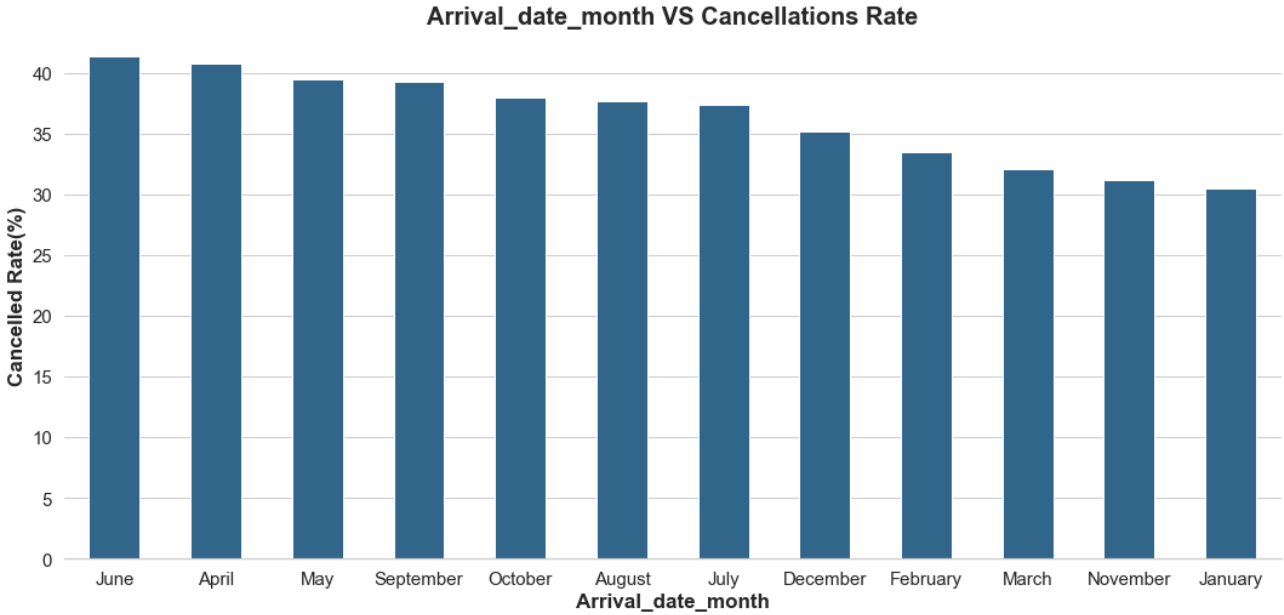
Peak Season is an important feature may affect Booking and Cancellation.

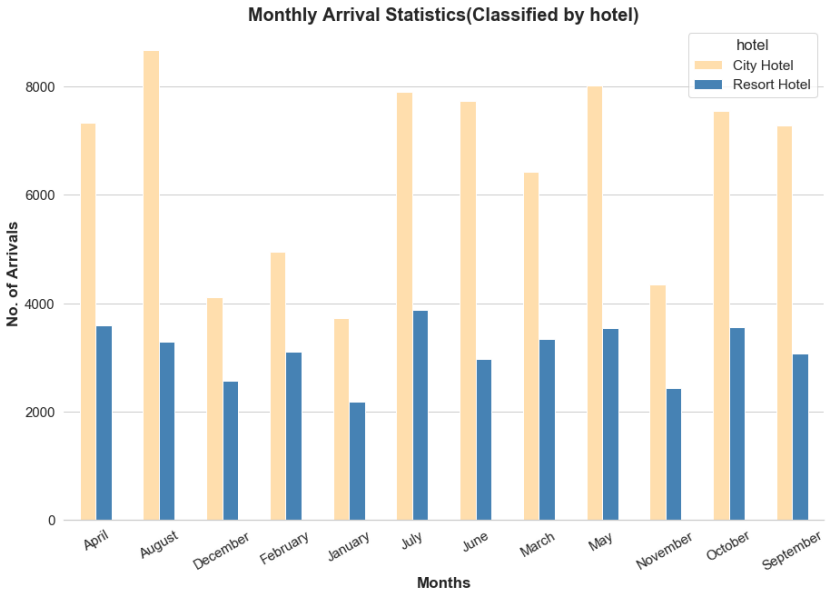
Bookings

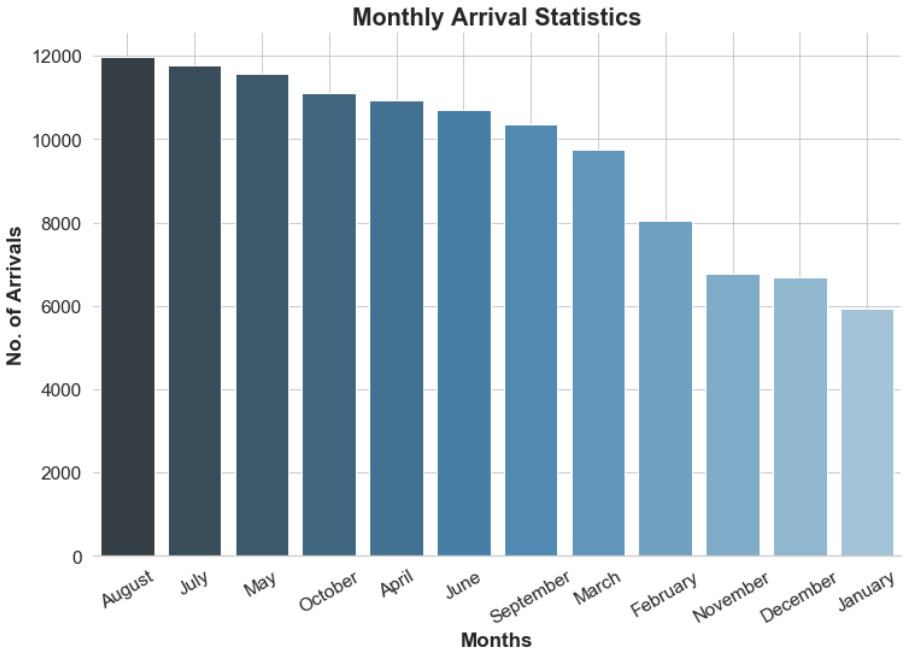
- Summer (August and July) are the most popular seasons for visitors Booking number in winter(Dec, Jan, Nov) is the lowest.

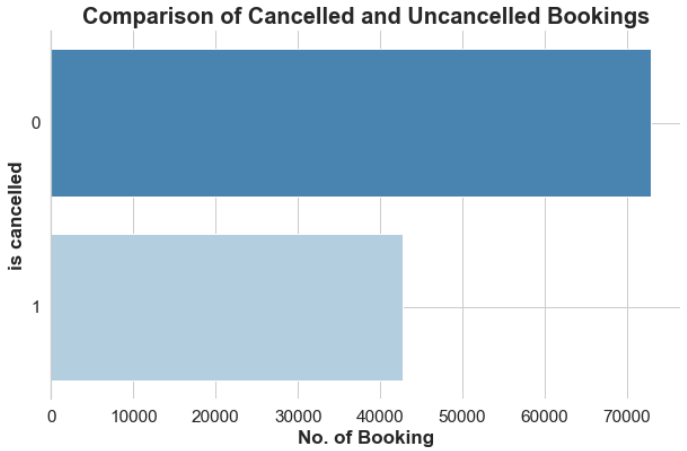
Cancellations

- Booking cancellation number is far less than unconcealed ones.
- The cancellation number is higher in April, May, June and July, which could be considered as the time before holiday.
- Bookings in April, May, June are more likely to be canceled.







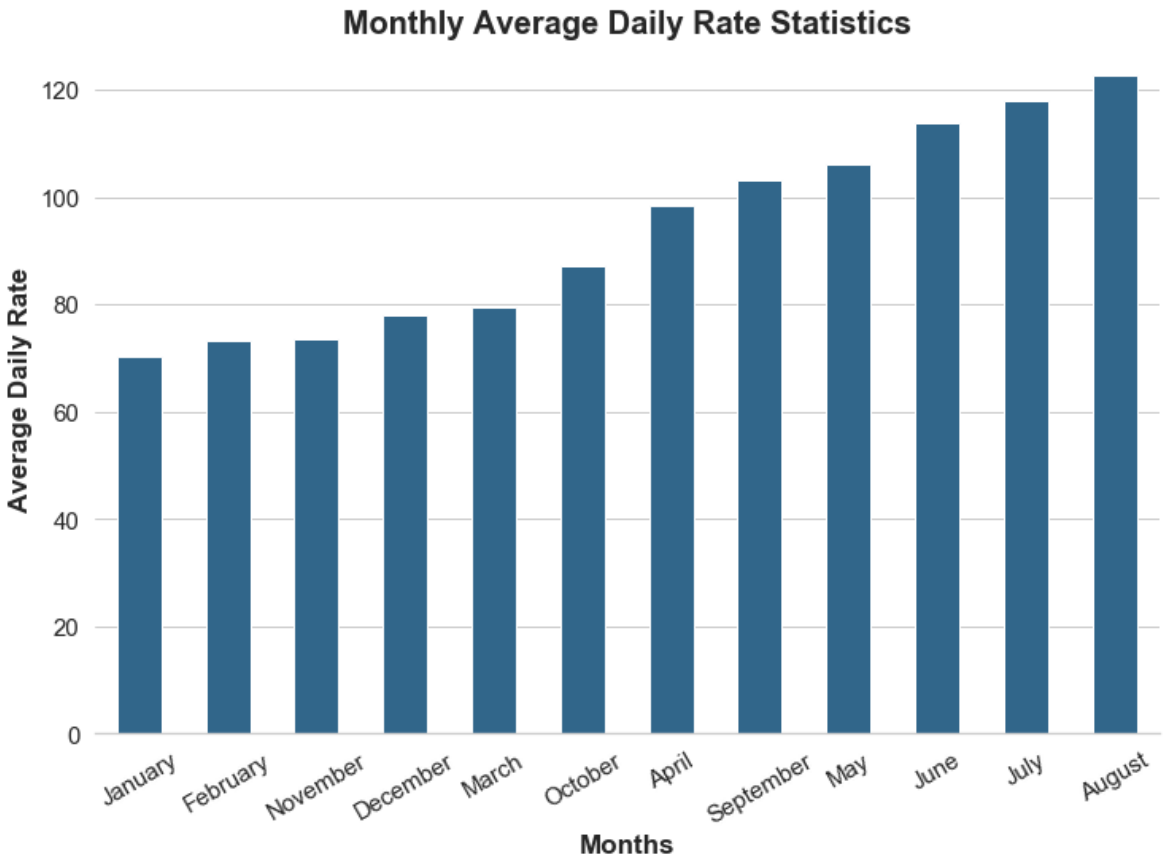


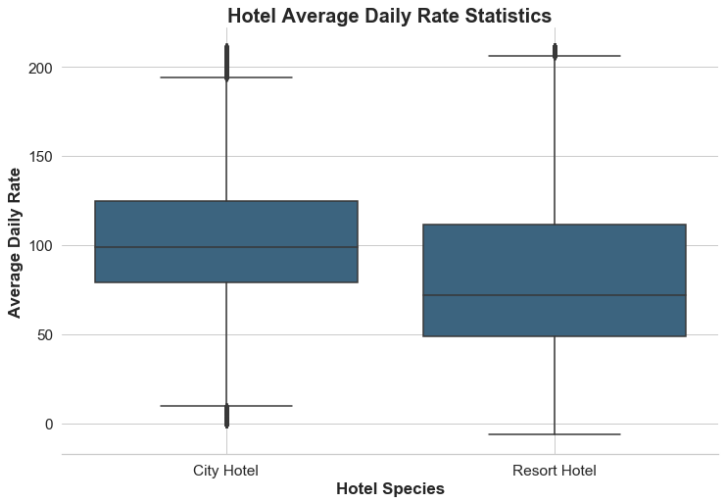
EDA- What is the price trend?

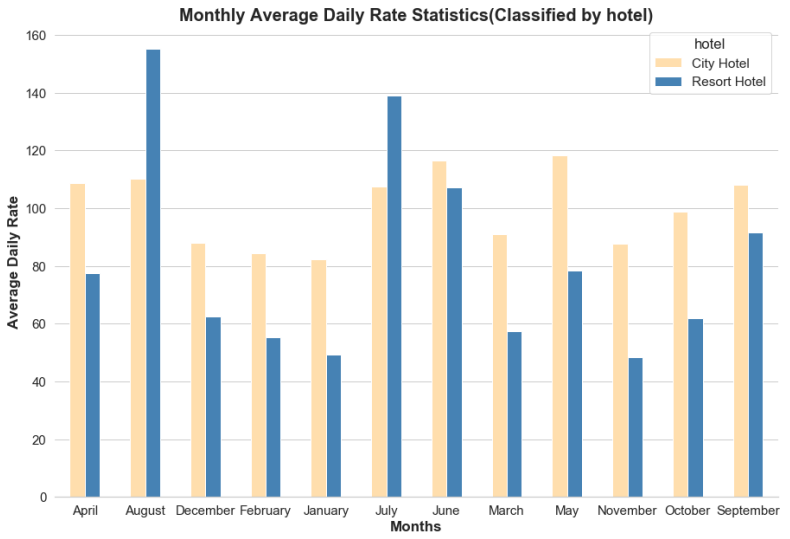
Prices are strongly influenced by the seasons.

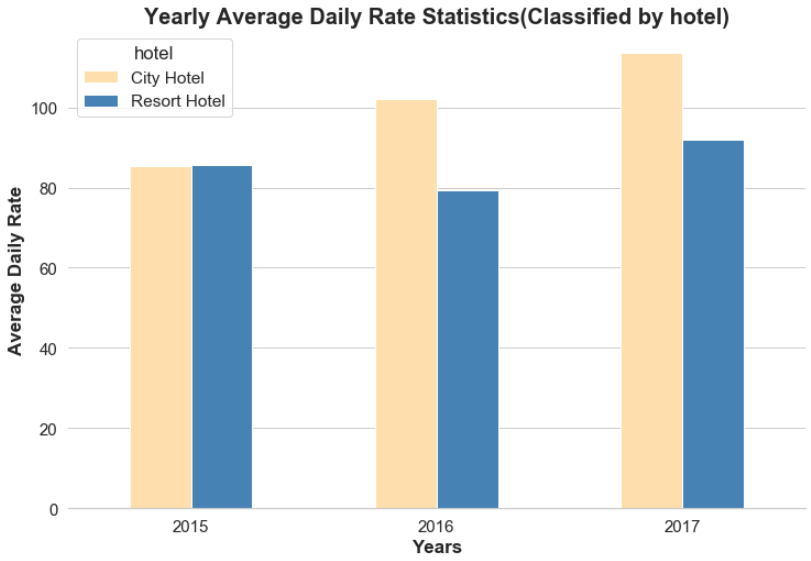
Bookings

- City hotels have a higher mean price than resort hotels.
- August has the highest adr. Summer has higher adr than winter.
- In summer resort hotels usually have a higher adr than city hotels. Resort hotels are greatly influenced by the seasons.
- Adr(Average Daily Rate) is growing year by year.







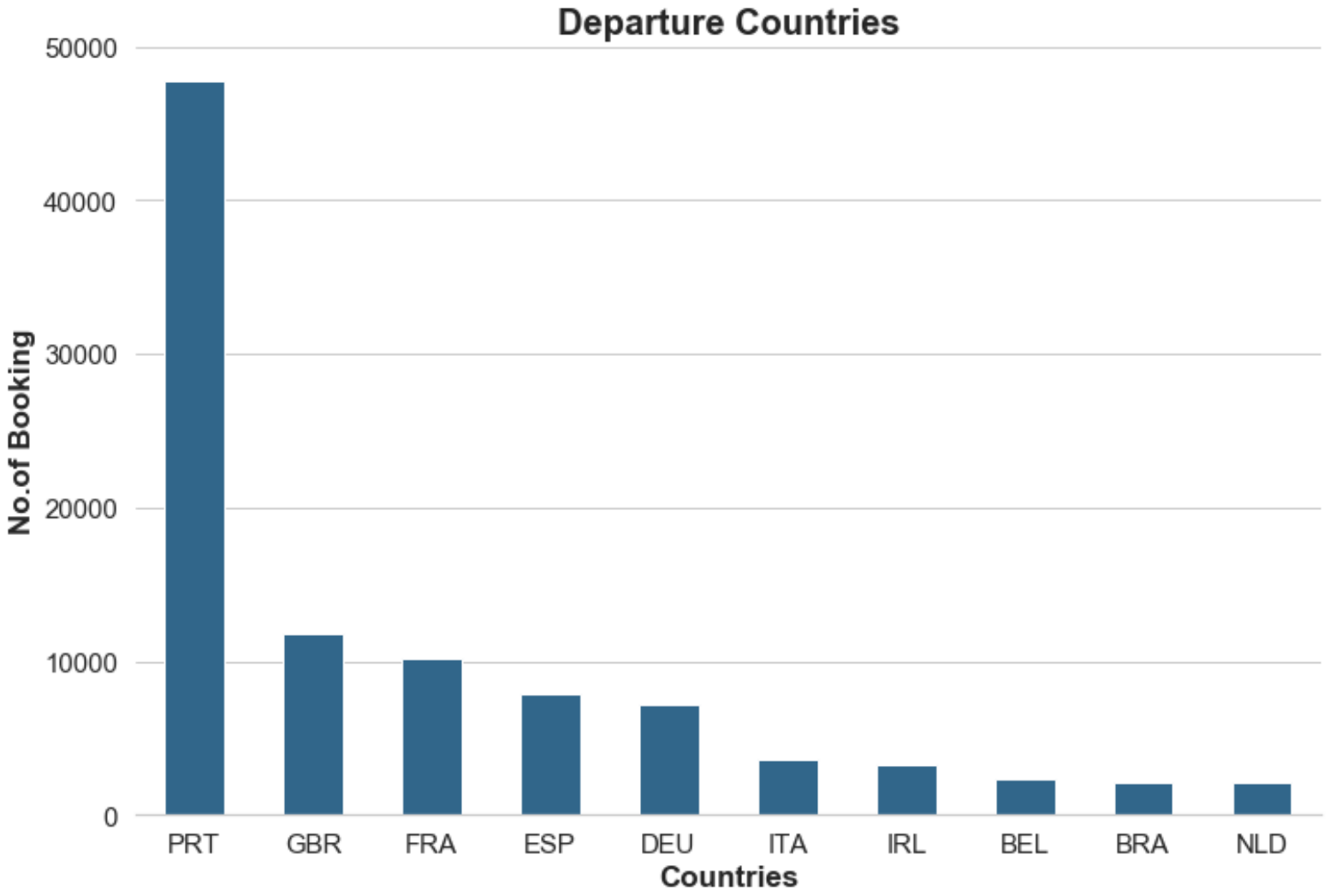


Springboard

EDA- Where are the visitors mainly from?

Most of customers are from **native.**

The dataset is created in Portugal. Except for Portugal, UK is the largest visitor's original country.



EDA- Which kind of hotel is more popular?

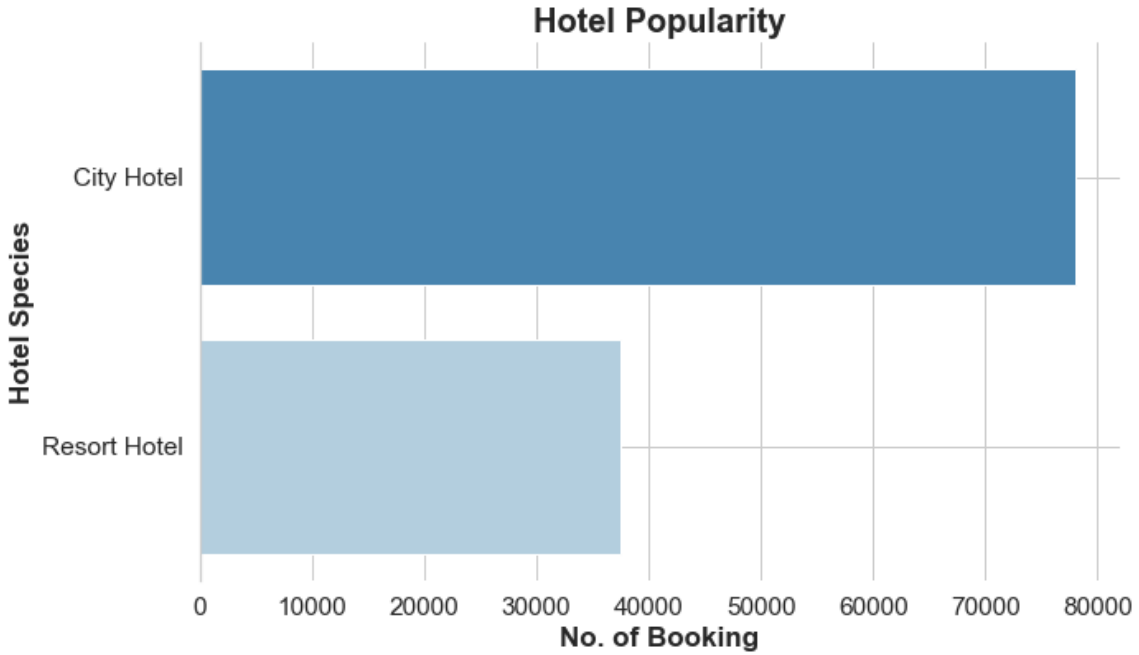
City hotels have a higher booking number as well as cancellations because of business trip.

Bookings

Booking number of city hotels is almost twice than resort hotels. That's because of a large proportion of business trip.

Cancellation

Combined with no canceled booking number, city hotels bookings were more likely to be canceled, due to the great proportion of business trips.



EDA- What kind of booking methods is more likely to be cancelled?

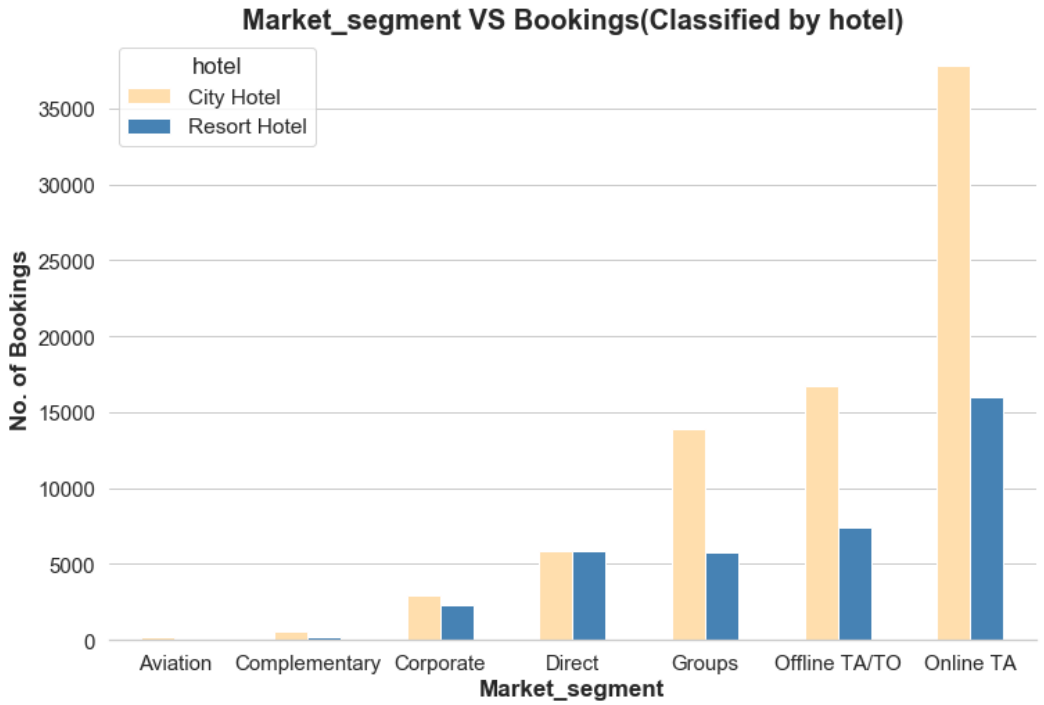
Most people prefer to choose the **Online Travel Agents**, which also has the largest cancellation number.

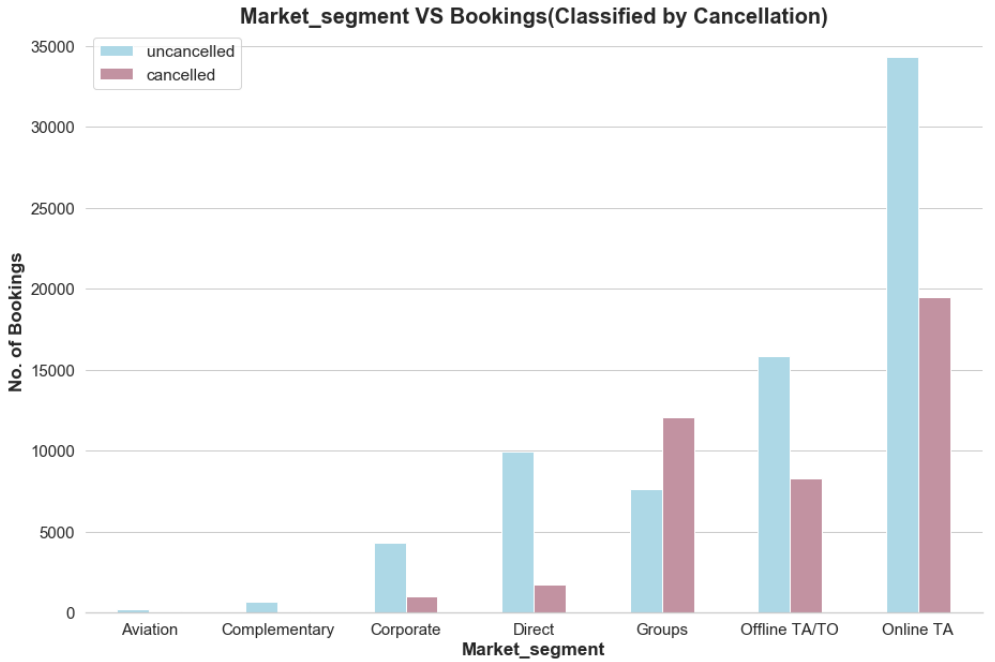
Bookings

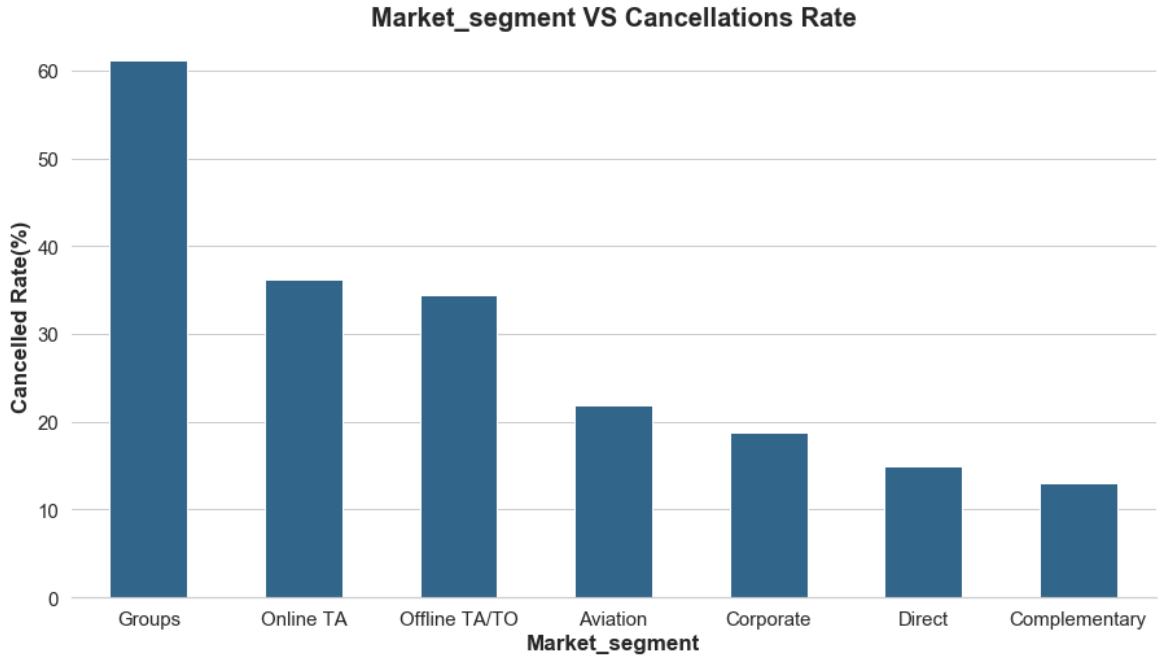
- Most people prefer to choose the Online Travel Agents.

Cancellation

- Group bookings are more likely to suffer cancellation.
- But Online TA has the largest cancellation number.







EDA- What the lead time may affect?

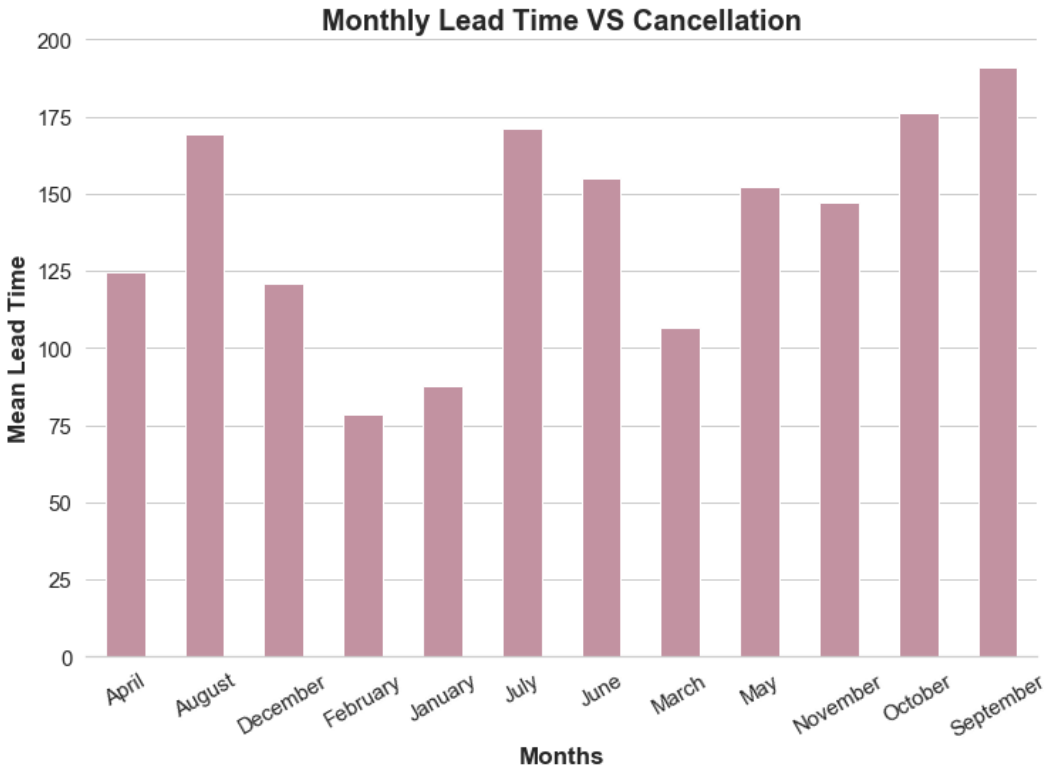
to some extent, as the **lead time** increases, the probability of cancellation increases.

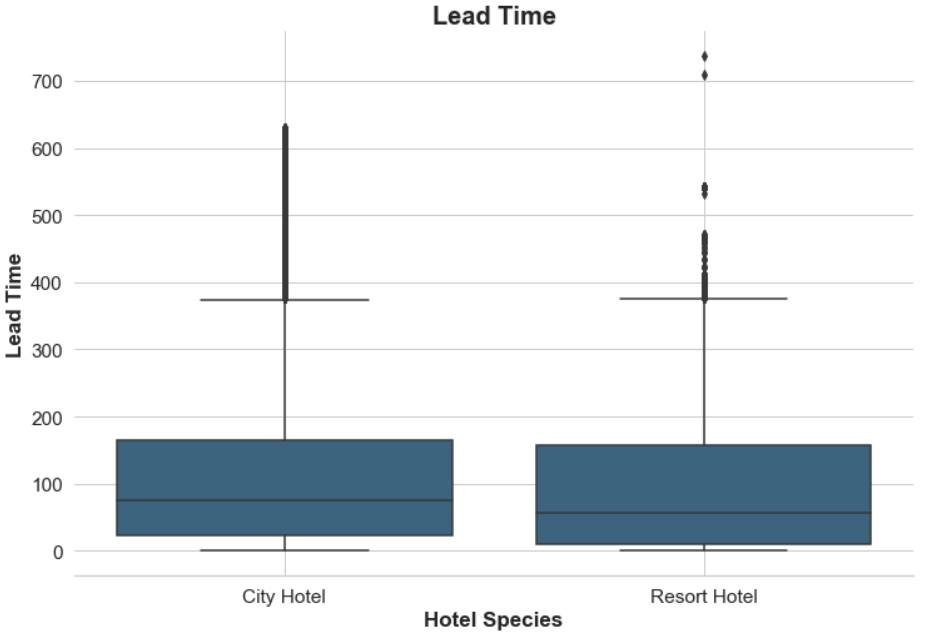
Bookings

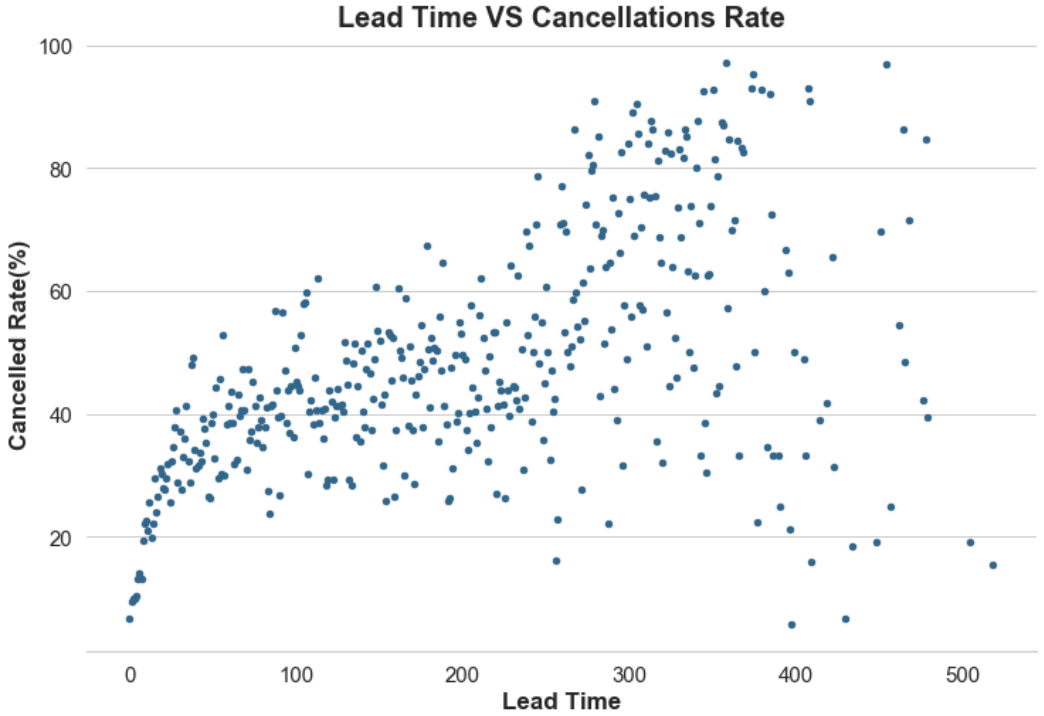
Most booking lead time concentrated distributes below 100 days. Resort hotels lead time is a little less than city hotels

Cancellation

- Group bookings are more likely to suffer cancellation.
- But Online TA has the largest cancellation number.





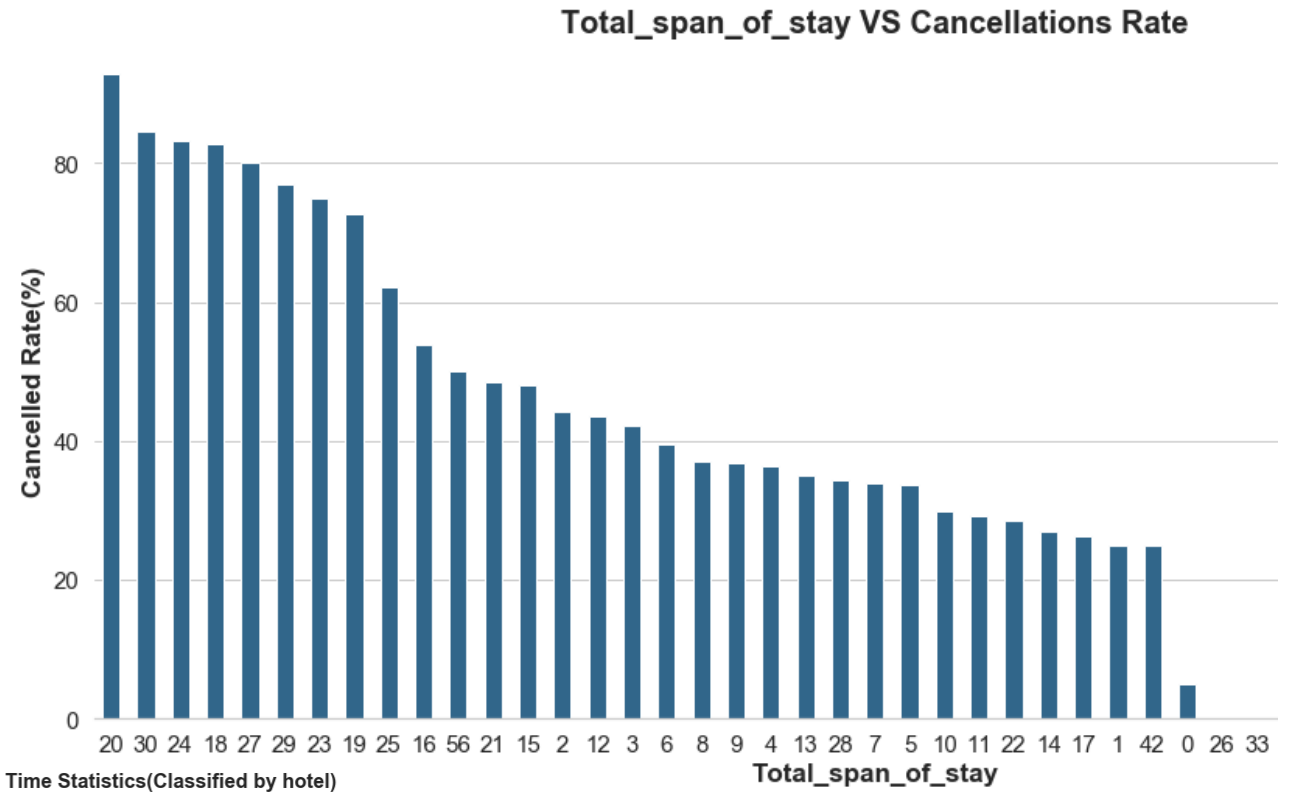


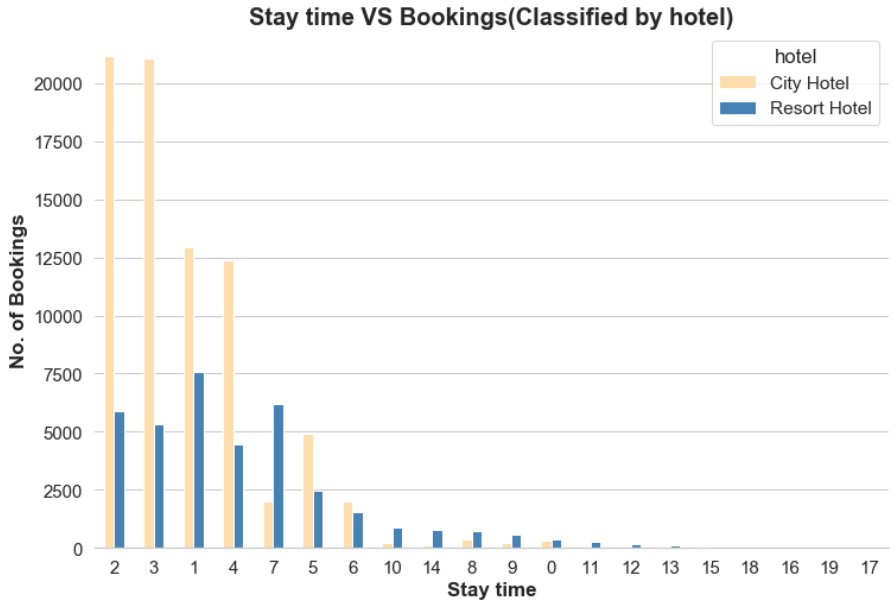
EDA- What the stay time may affect?

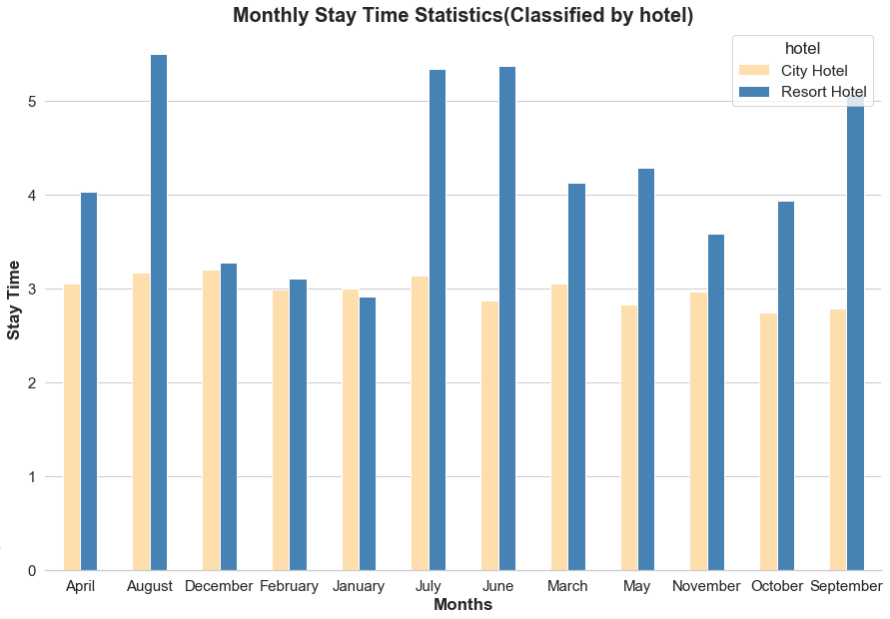
Stay time is determined by season and may affect the cancellation rate.

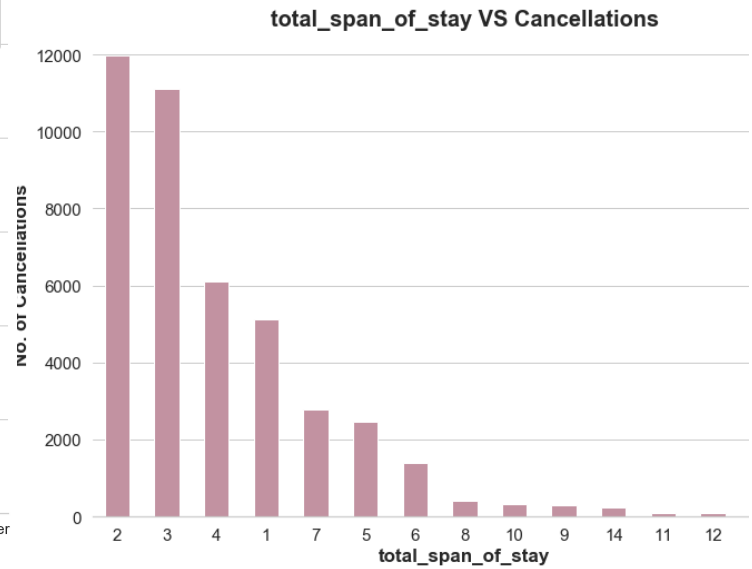
Bookings
- Most bookings have a 1-3 days stay time.
- Resort hotels stay time is a little longer than city hotels.
- Resort hotels usually have a longer stay time in summer.

Cancellation
- 20-30 days stay time has the largest cancellation rate.
- But 2-4 days stay time has the largest cancellation number.







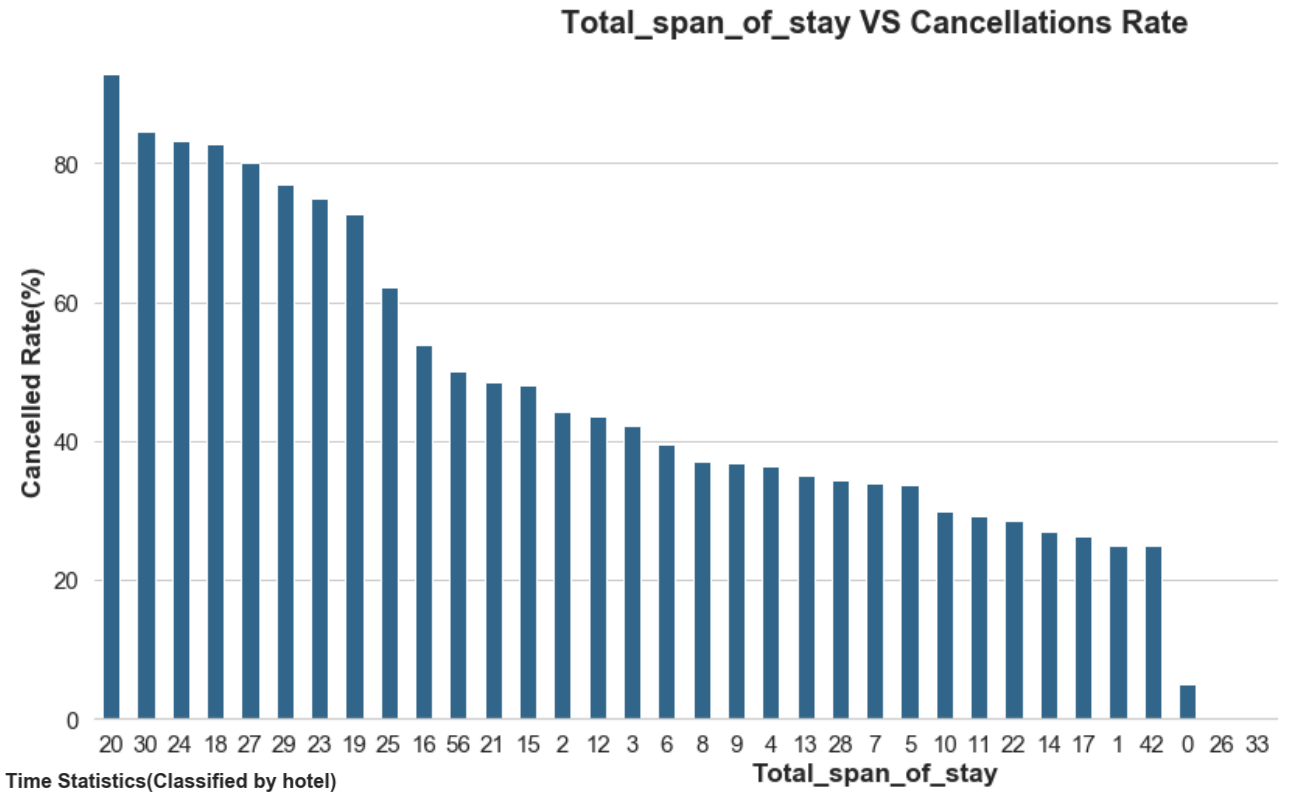


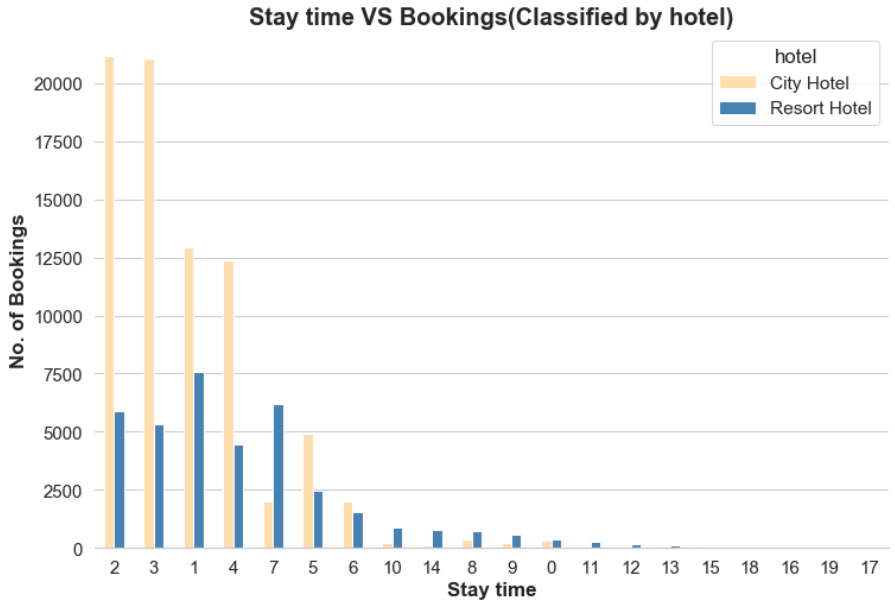
EDA- What the stay time may affect?

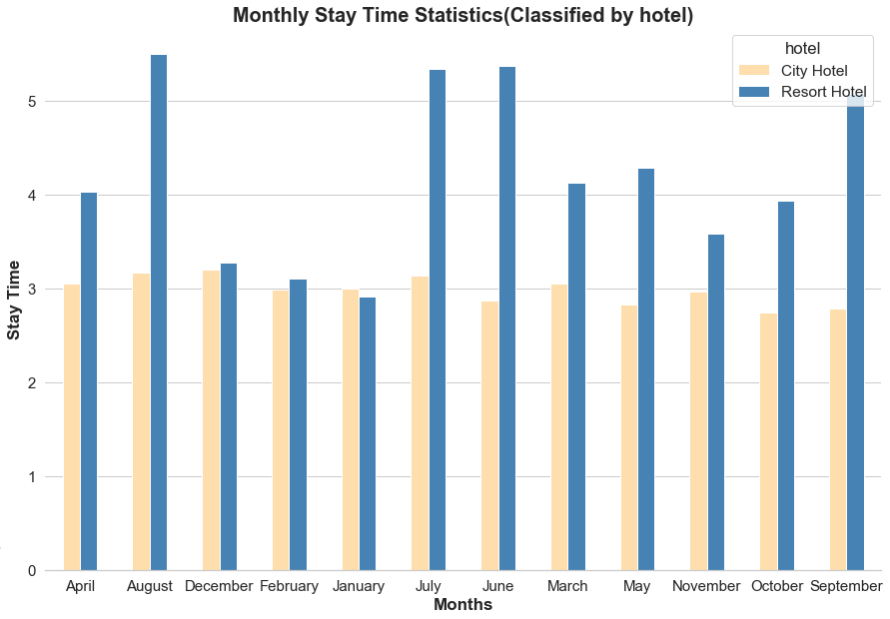
Stay time is determined by season and may affect the cancellation rate.

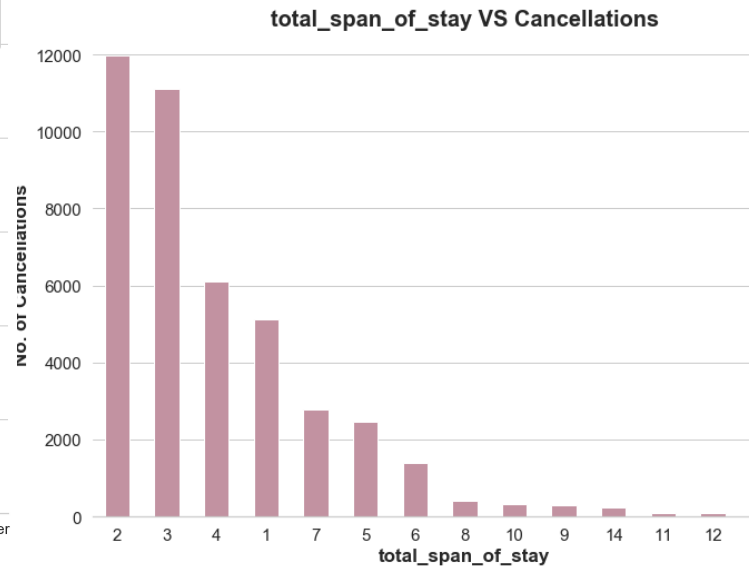
Bookings
- Most bookings have a 1-3 days stay time.
- Resort hotels stay time is a little longer than city hotels.
- Resort hotels usually have a longer stay time in summer.

Cancellation
- 20-30 days stay time has the largest cancellation rate.
- But 2-4 days stay time has the largest cancellation number.





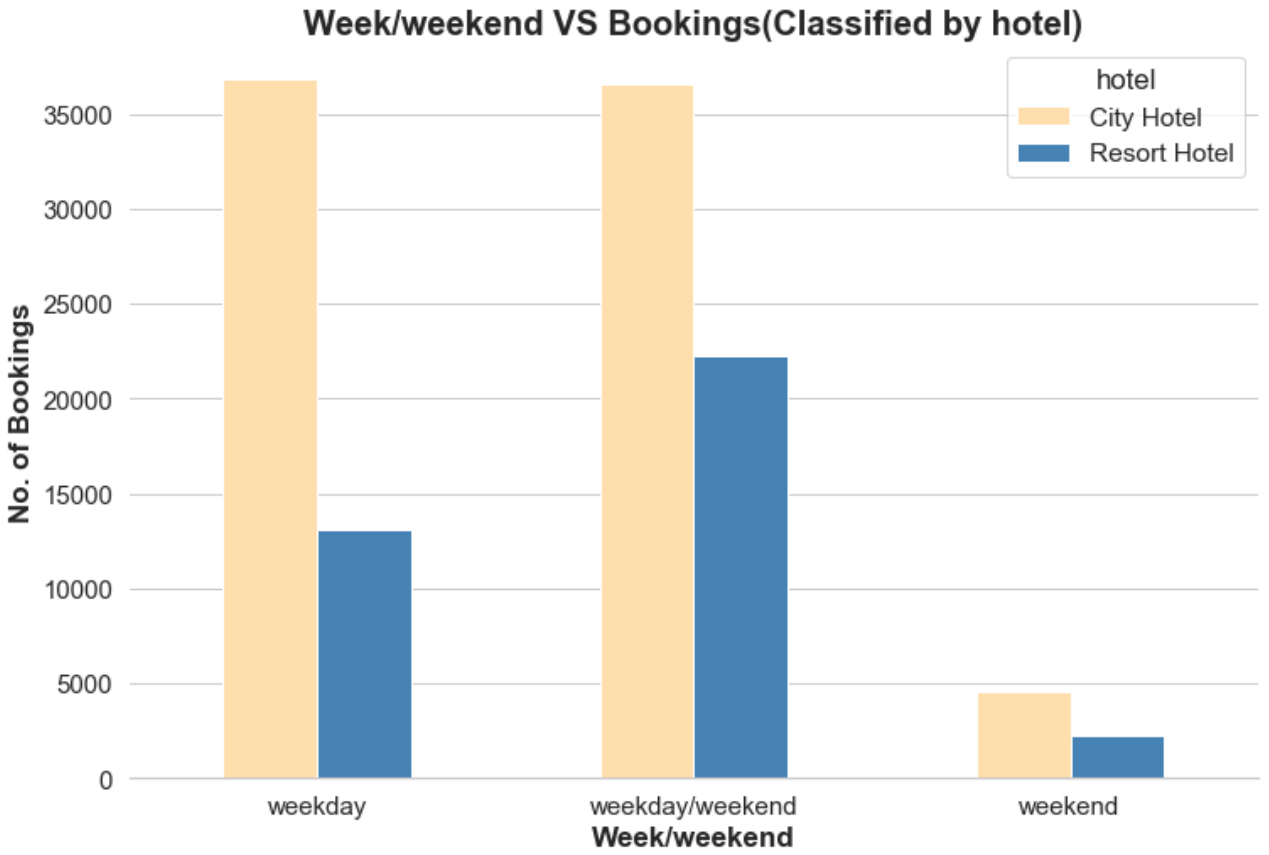


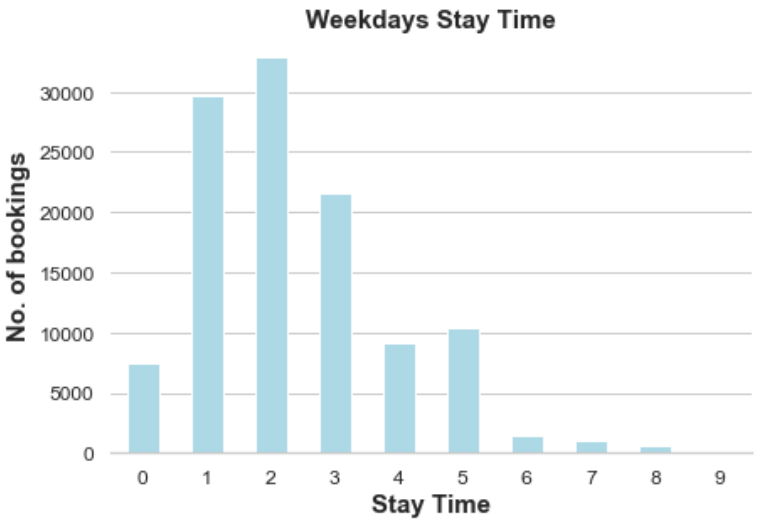


EDA- People prefer to book on weekday or weekend?

Most bookings happens on weekdays.

- Booking on weekdays is far more than weekends days.
- Stay time on weekdays is longer than weekends.
- Especially for city hotels, because of the business trip.

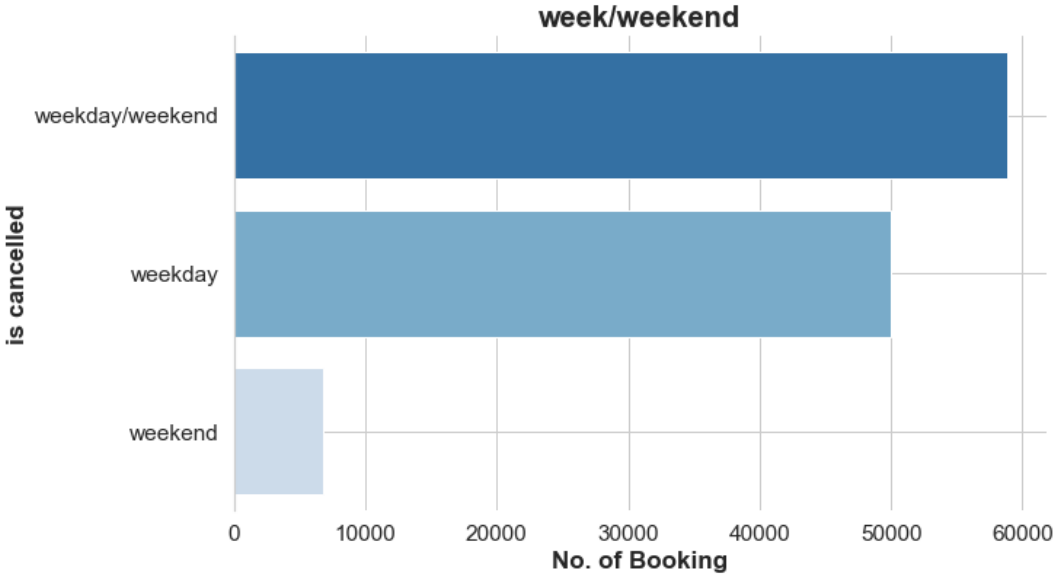


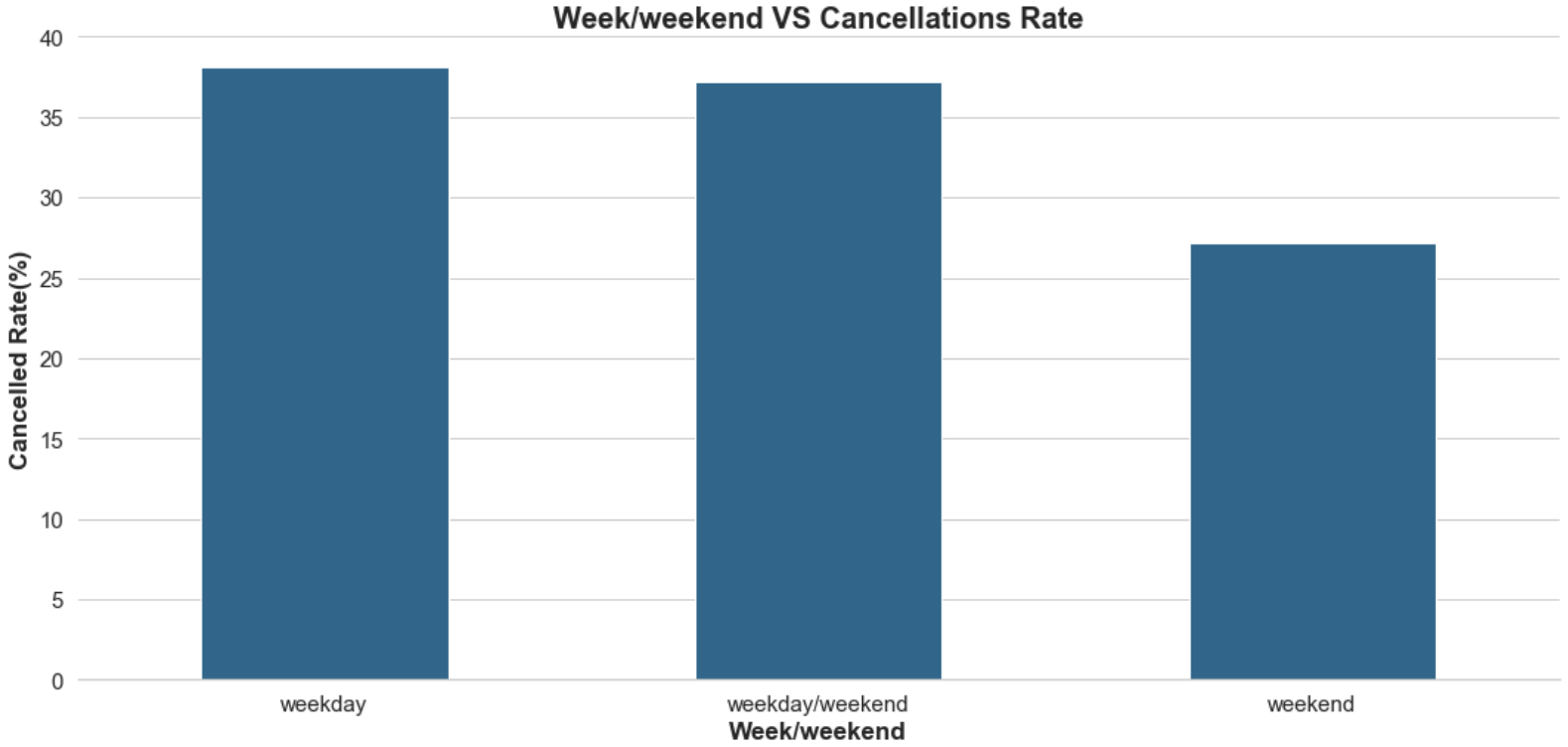


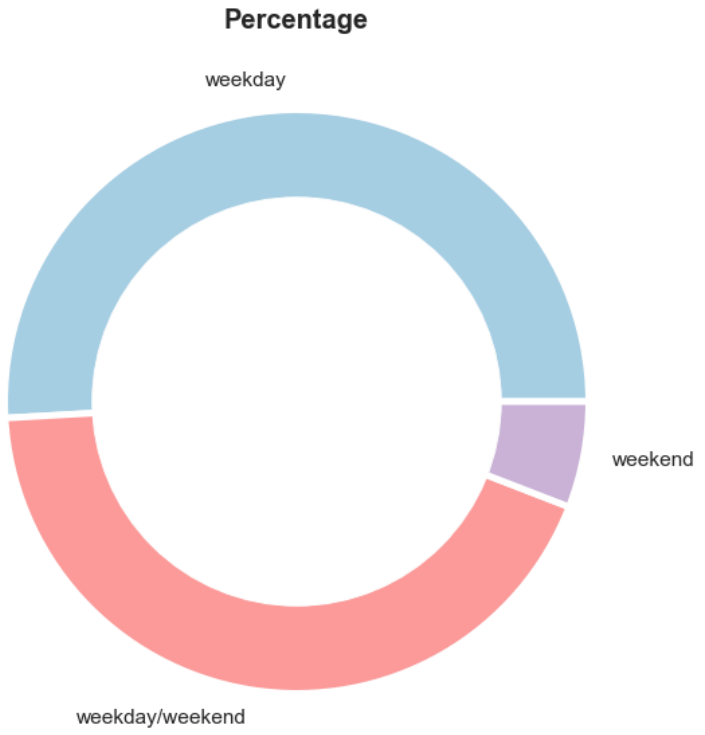
EDA- People prefer to cancel on weekday or weekend?

Most cancellations happens on weekdays.

- Weekend has the smallest cancellations.
- People are more likely to cancel booking on weekdays.



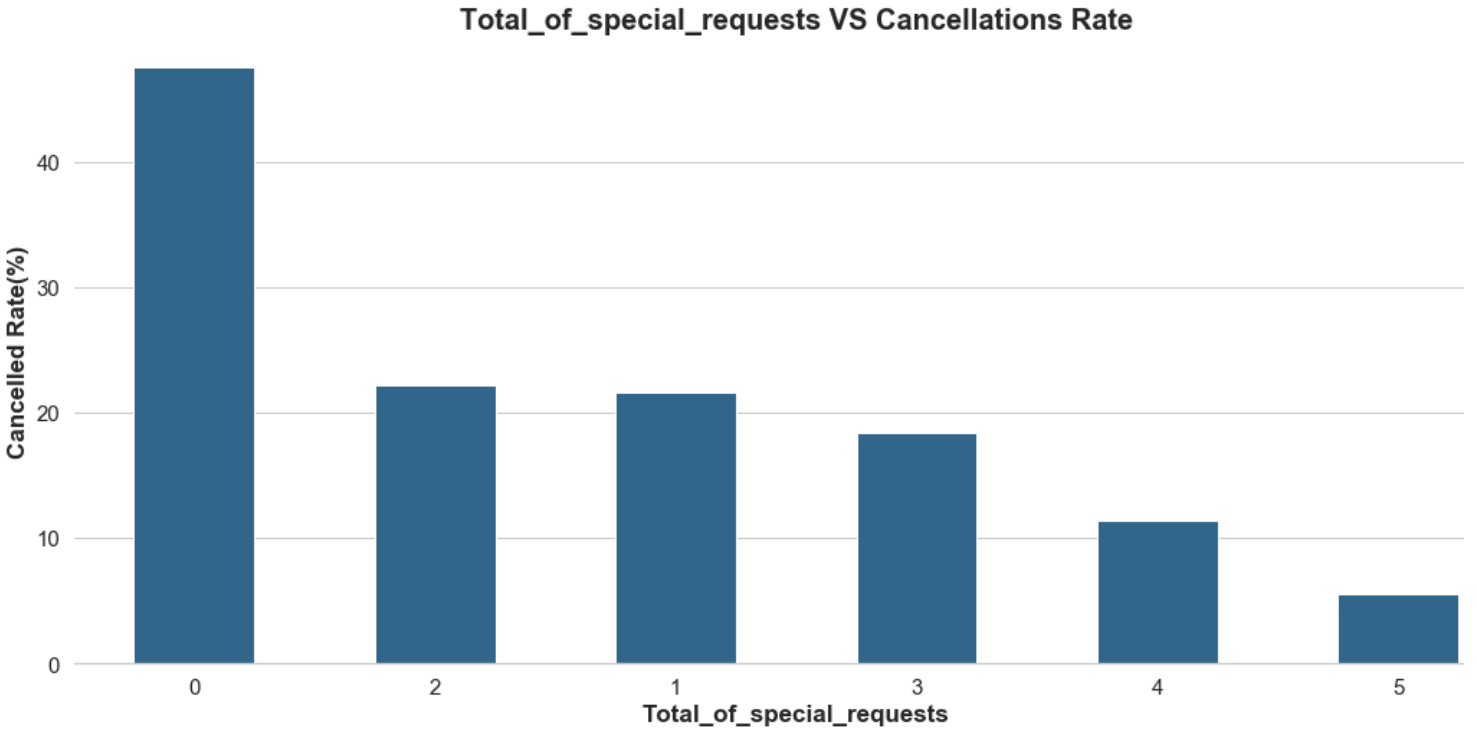




EDA- What does special requests mean to booking and cancellation?

Special requests usually affect the cancellation.

- Most bookings had no special request.
- It is obvious that people without any request had the largest cancellation rate. As the number of special requests increases, the rate comes down.



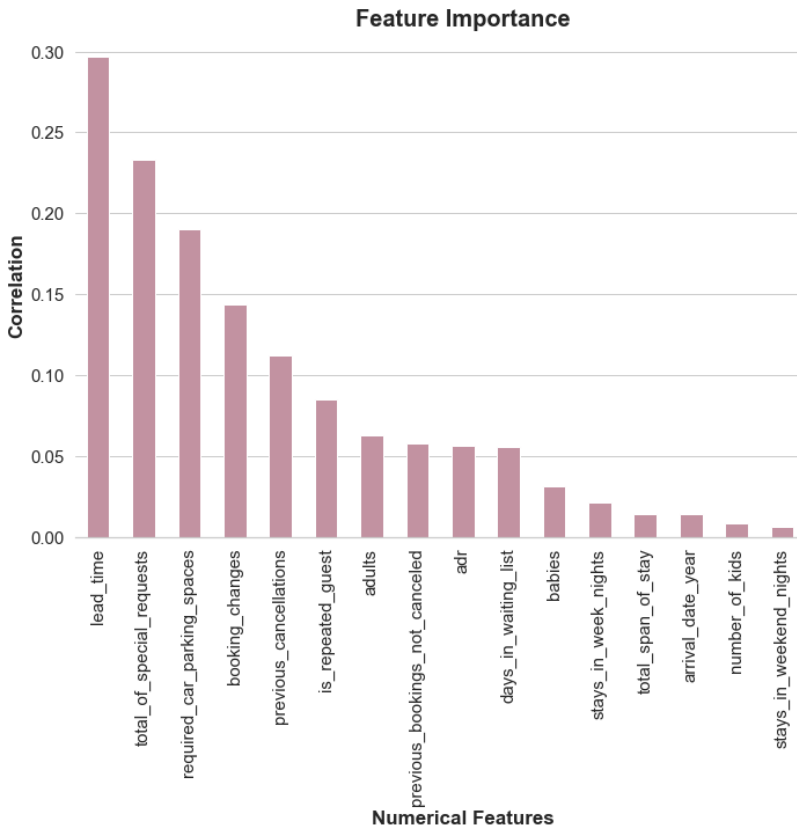
STATISTICAL ANALYSIS

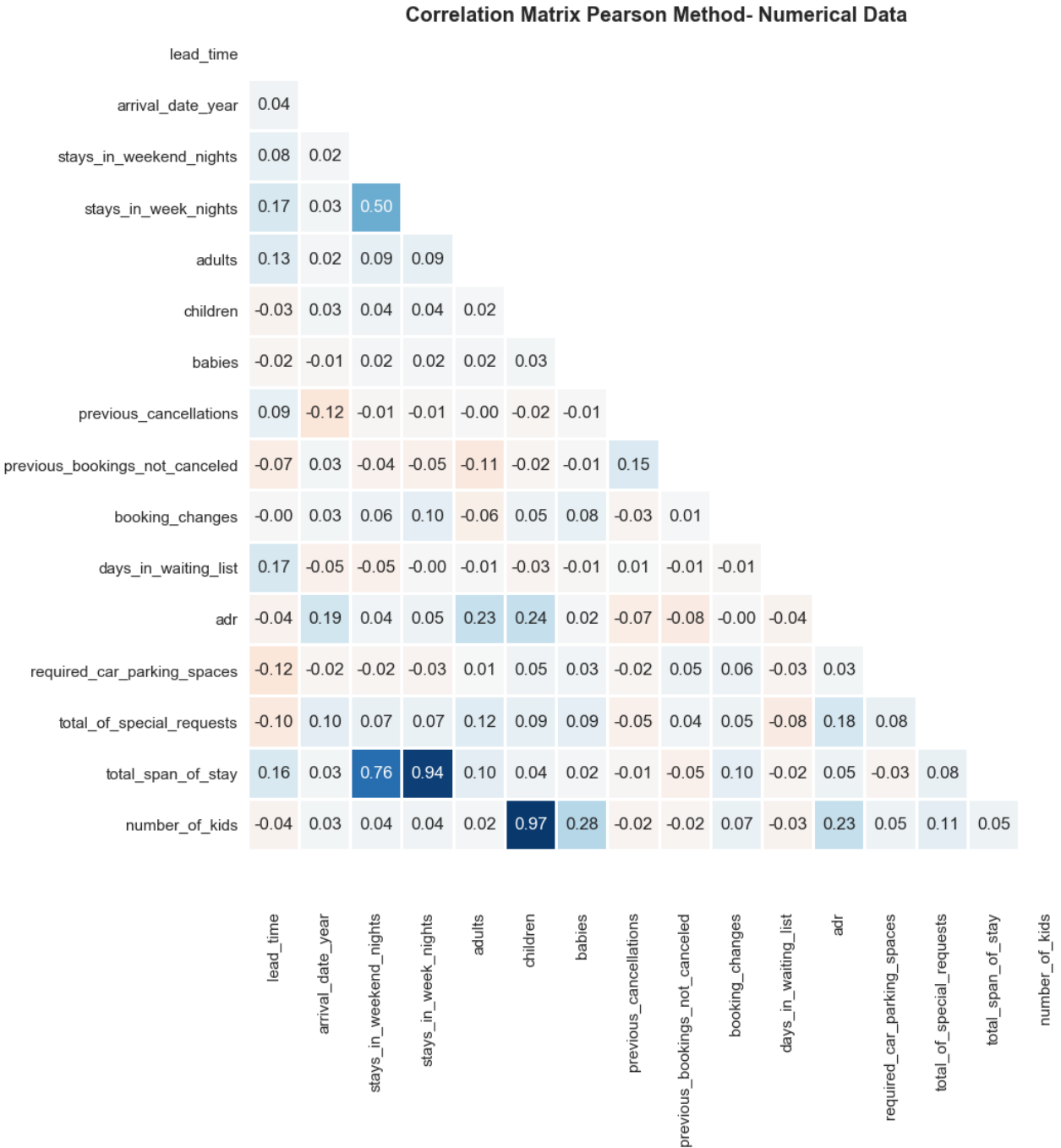
Evaluate the correlation of all the features

- ‘lead_time’, ‘total_of_special_requests’, ‘required_car_parking_spaces’, ‘booking_changes’ and ‘previous_cancellations’ are the 5 most important numerical features.
- ‘booking_change’ Will be affected by target variables, so I won't include it.

lead_time	0.296587
total_of_special_requests	0.233214
required_car_parking_spaces	0.190400
booking_changes	0.143636
previous_cancellations	0.111872
is_repeated_guest	0.084991
adults	0.062545
previous_bookings_not_canceled	0.057877
adr	0.056329
days_in_waiting_list	0.055710
babies	0.030956
stays_in_week_nights	0.021369
total_span_of_stay	0.013695
arrival_date_year	0.013650
number_of_kids	0.008493
stays_in_weekend_nights	0.005782
children	0.000444

Name: is_canceled, dtype: float64

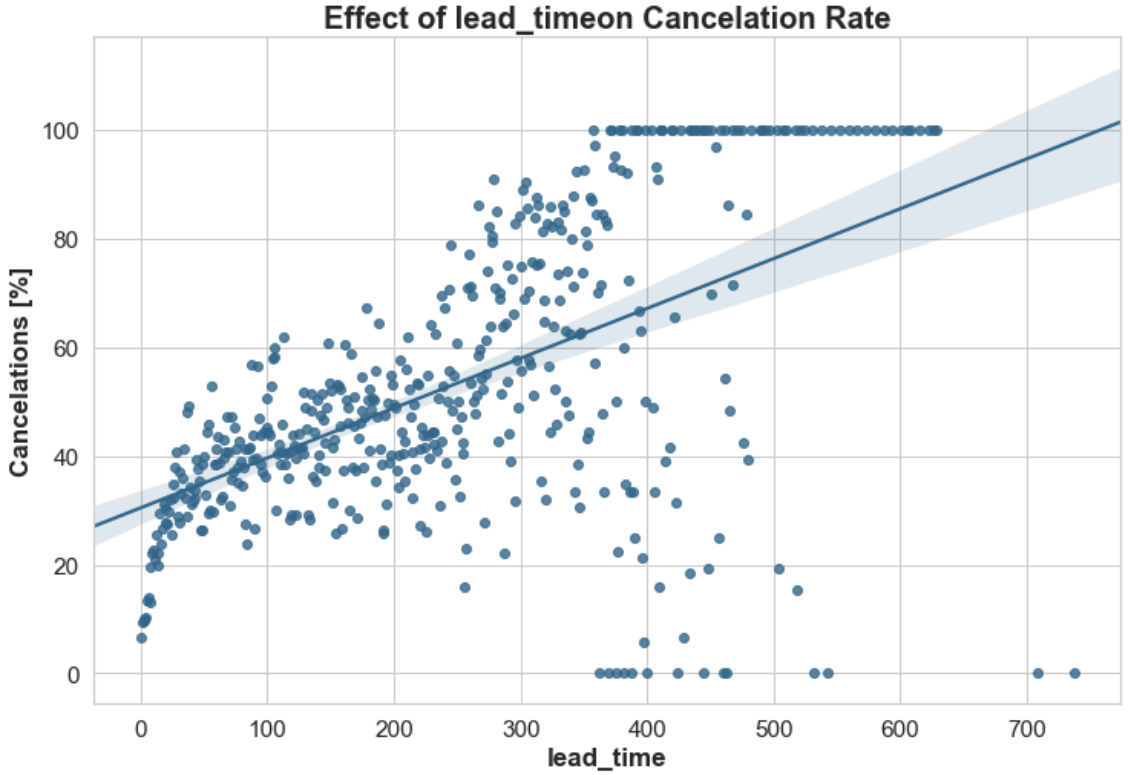


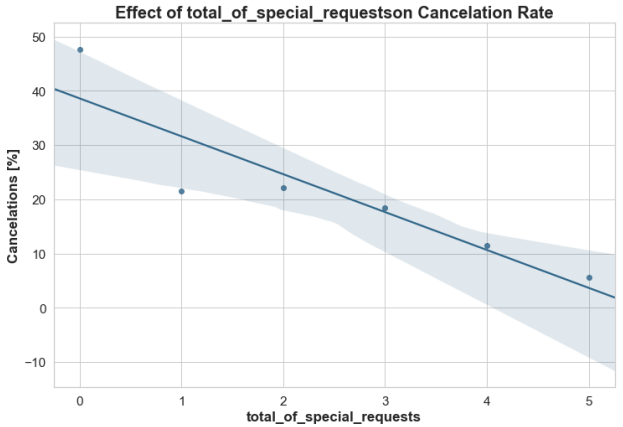


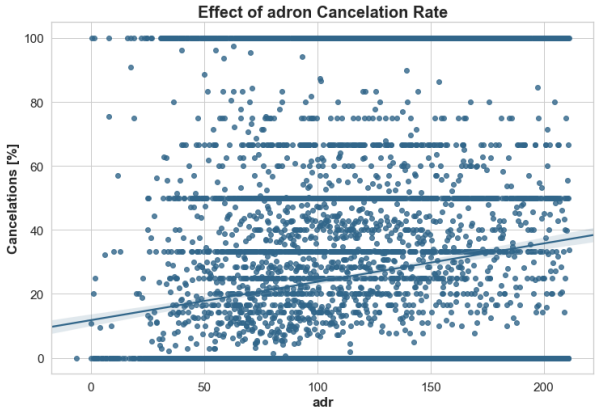
STATISTICAL ANALYSIS

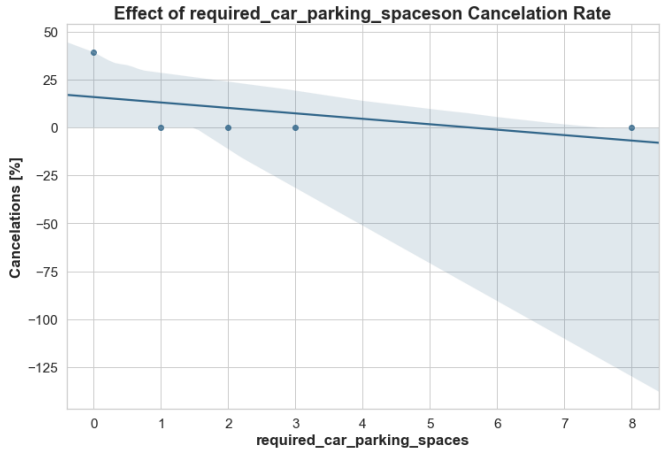
Research the most important numerical features

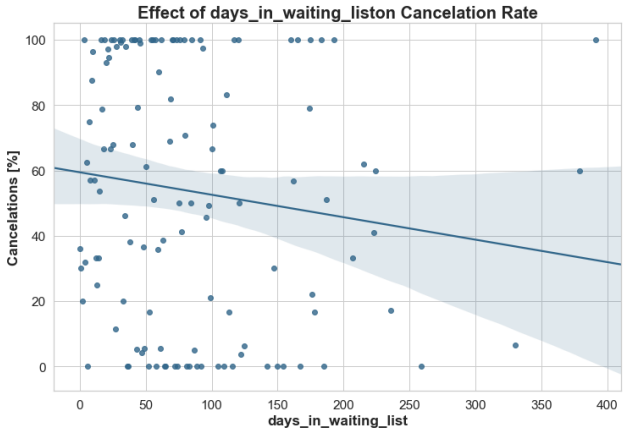
- After researching the scatter diagrams, none of the features had a linear correlation with the cancellation rate.
- However, I could find out some law between the lead time and cancellation rate. It seems that before 50 days lead time the cancellation raised quite fast. After that it is obviously slowed down.





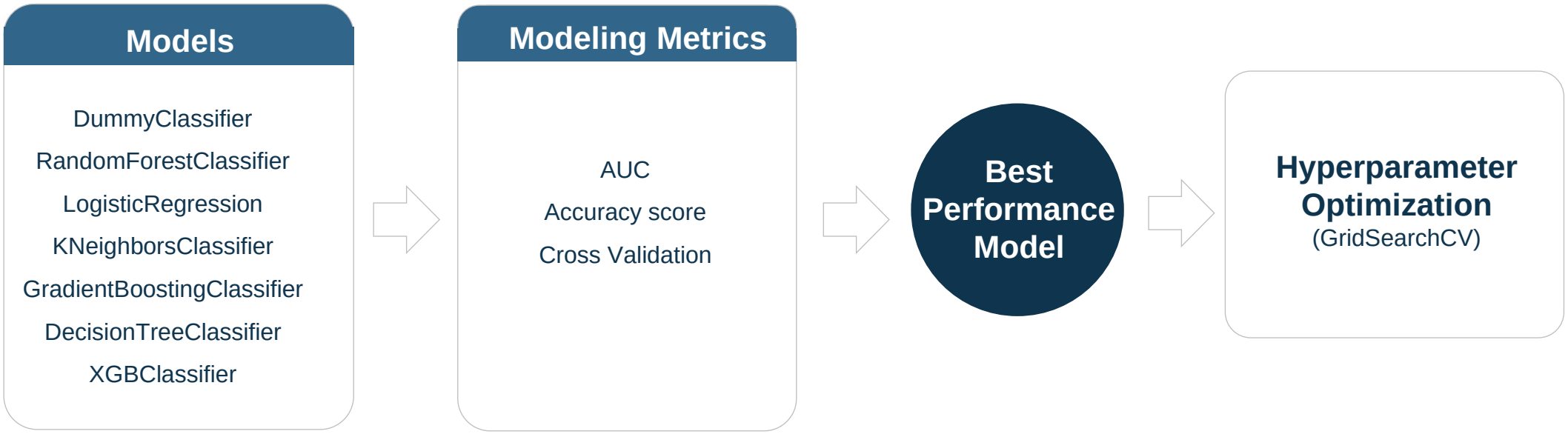






Modeling
idea

Modeling tools
- Sklearn
- Py-xgboost

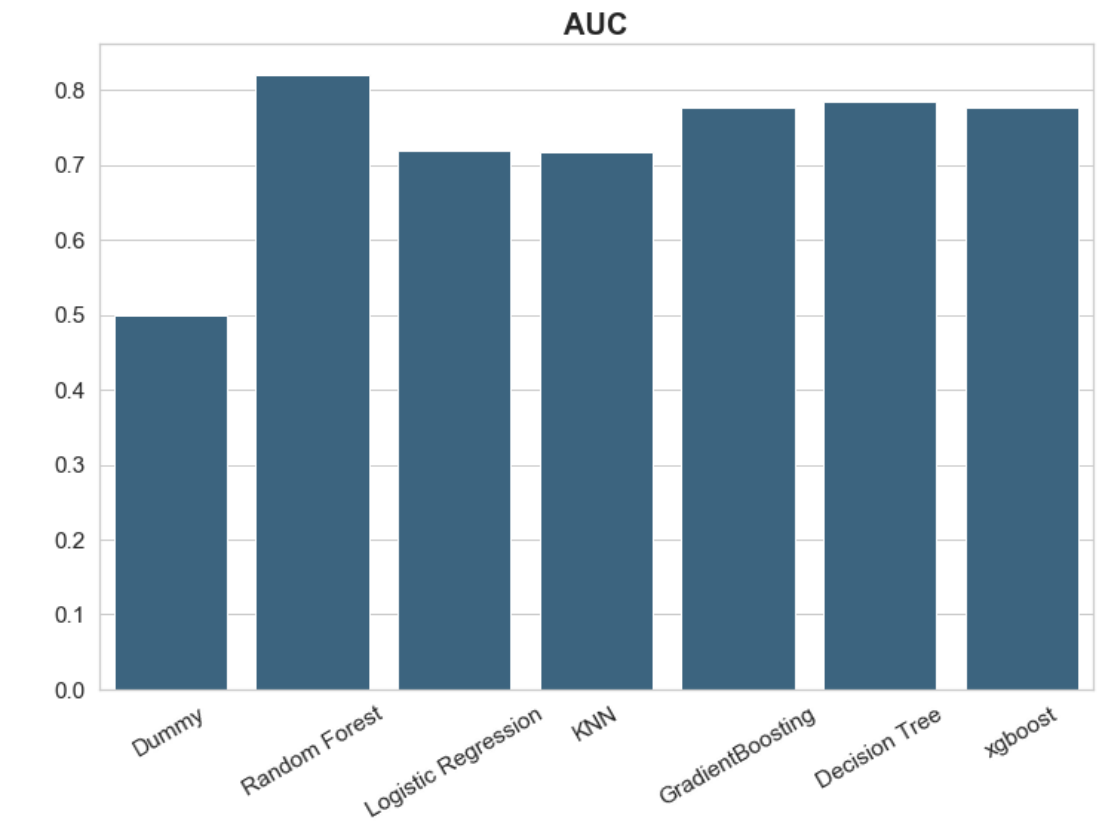


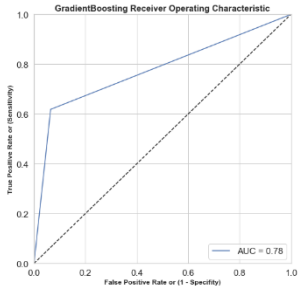
Modeling

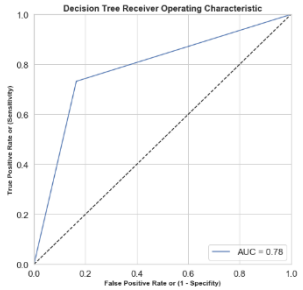
AUC

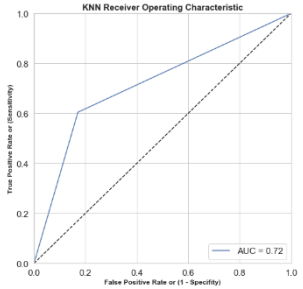
RandomForest got the highest AUC score, which is **0.82.**

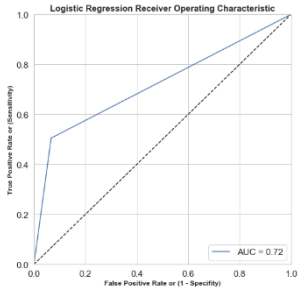
	Dummy	Random Forest	Logistic Regression	KNN	GradientBoosting	Decision Tree	xgboost
0	0.498862	0.820246	0.719212	0.716848	0.7769	0.785097	0.775478

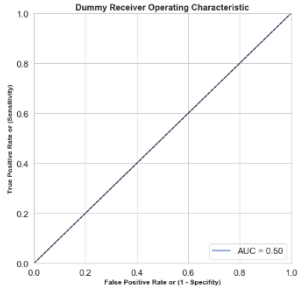


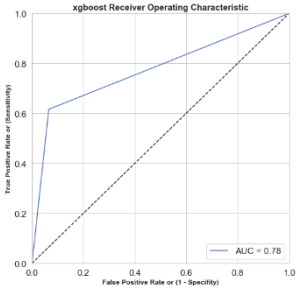


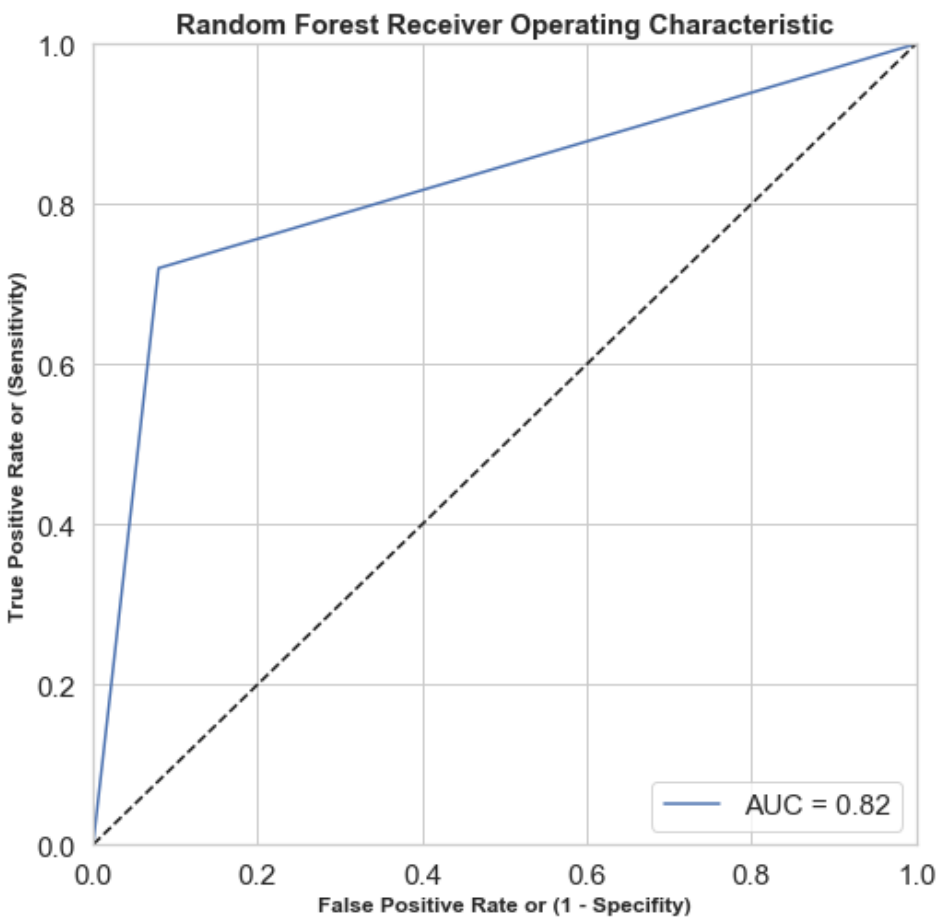










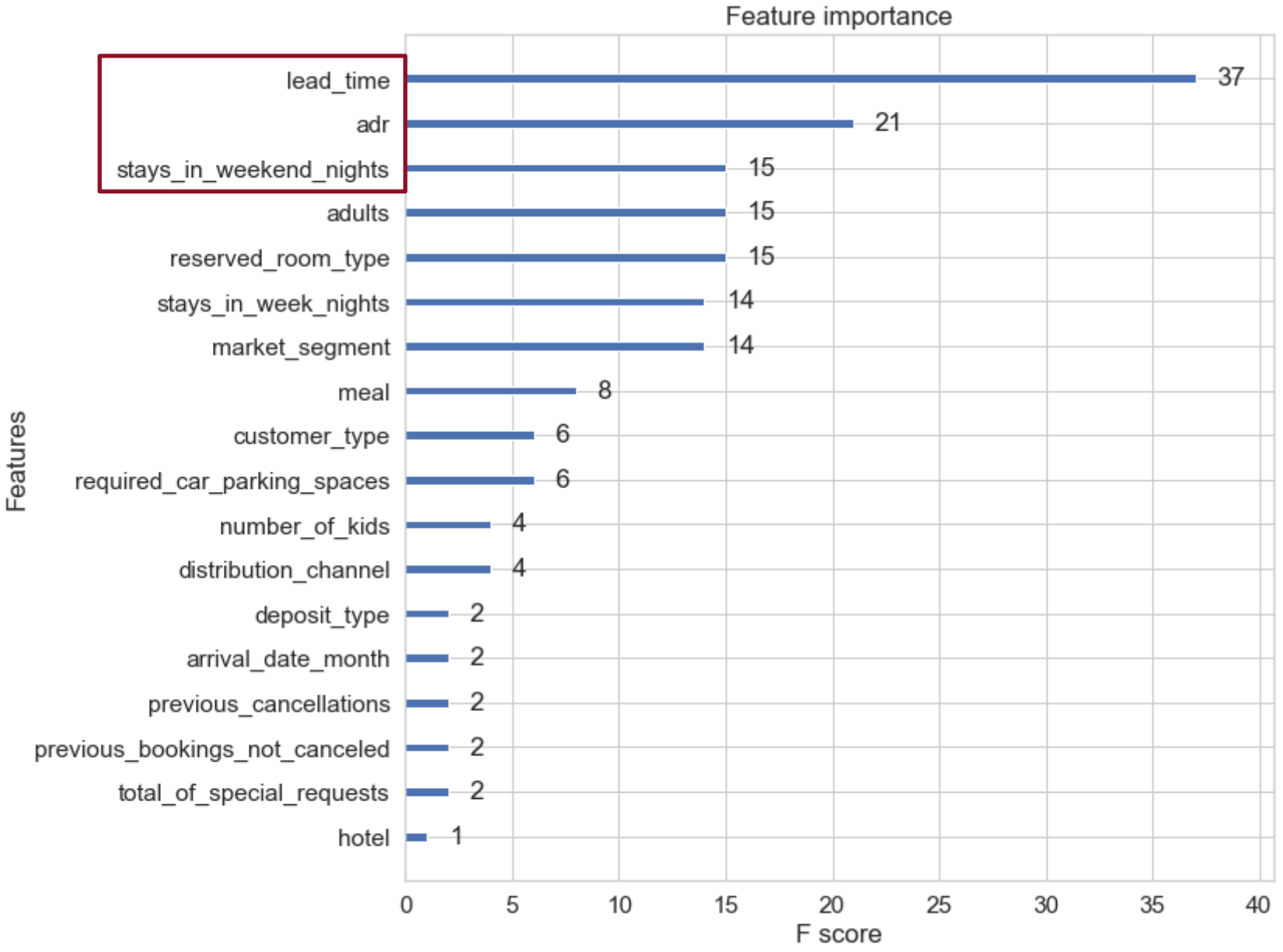


Modeling

Xgboost Features Importance

I test the Xgboost Feature importance to check the overfit.

If some feature is particularly more important than others, it means the risk of overfitting



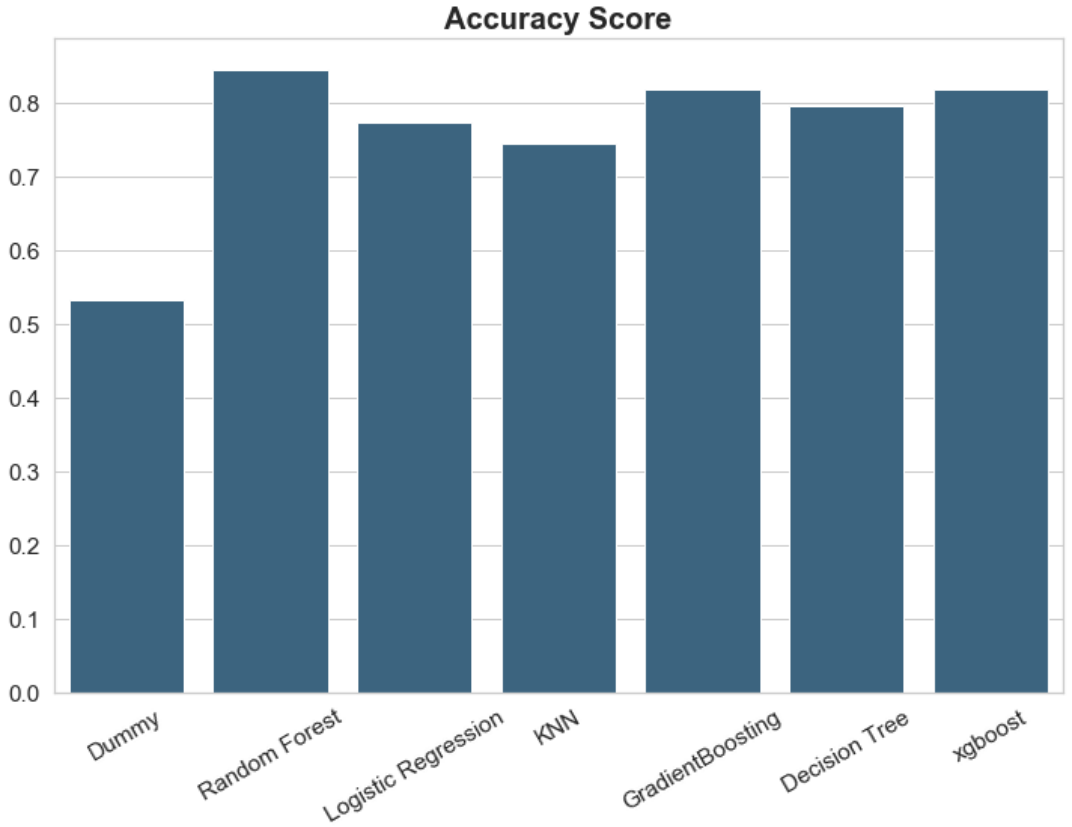
Modeling

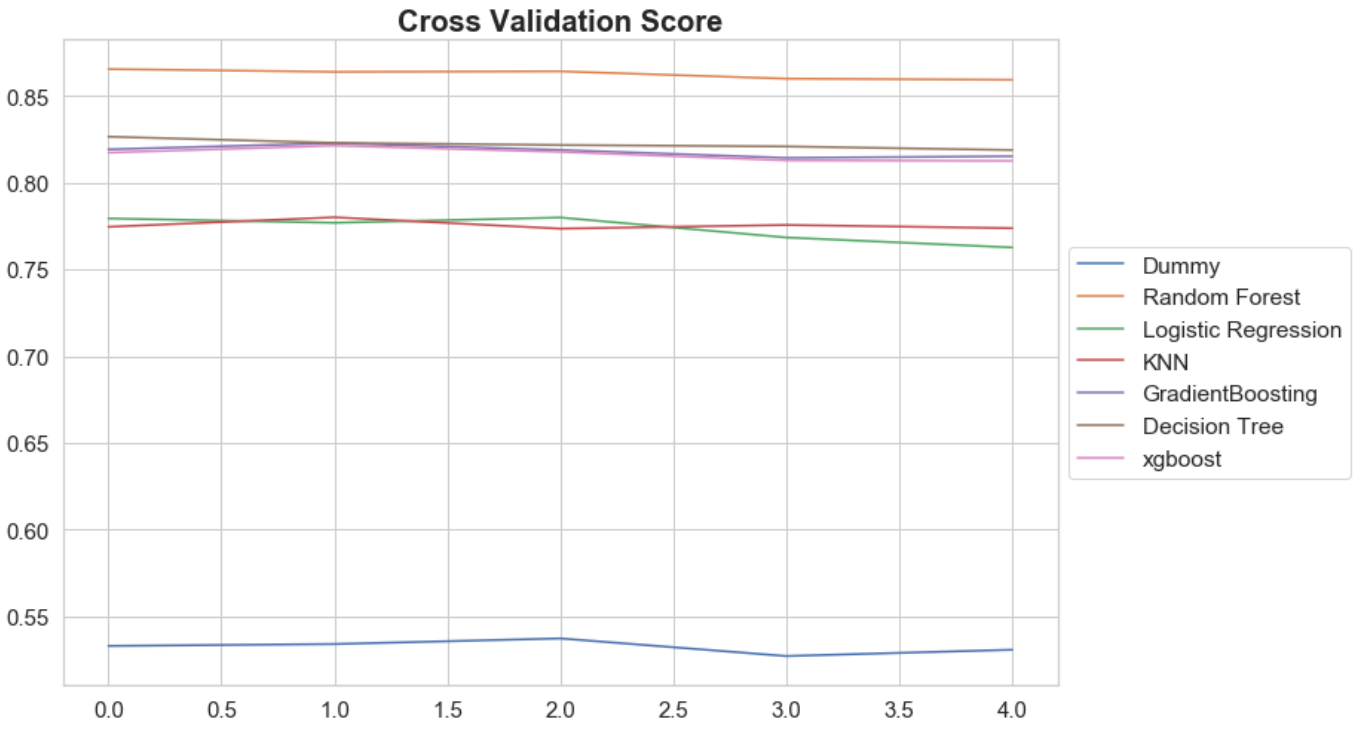
Accuracy score & Cross Validation

Random Forest had the greatest performance in both Accuracy Score and Cross Validation

	Dummy	Random Forest	Logistic Regression	KNN	GradientBoosting	Decision Tree	xgboost
0	0.533068	0.865522	0.779445	0.774601	0.819240	0.826550	0.817336

	Dummy	Random Forest	Logistic Regression	KNN	GradientBoosting	Decision Tree	xgboost
count	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000
mean	0.532540	0.862598	0.773502	0.775553	0.818106	0.822215	0.816402
std	0.003790	0.002757	0.007597	0.002669	0.003391	0.002855	0.003689
min	0.527229	0.859336	0.762706	0.773563	0.814352	0.818850	0.812622
25%	0.530862	0.859985	0.768459	0.773822	0.815260	0.820927	0.812881
50%	0.533068	0.863921	0.776937	0.774601	0.818894	0.821748	0.817336
75%	0.534149	0.864224	0.779445	0.775682	0.819240	0.823003	0.817769
max	0.537393	0.865522	0.779965	0.780094	0.822786	0.826550	0.821402





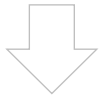
Modeling

Random Forest Optimization

There is a little improvement of hyperparameter optimization.
The important features of Random Forest is different form Xgboost.

Hyperparameter Optimization

auc score: 0.8202
Accuracy_score: 0.8455
Cross Validation Score: 0.8603



auc score: 0.8204
Accuracy_score: 0.8460
Cross Validation Score: 0.8605

Random Forest Feature Importance

1)	lead_time	0.201259
2)	adr	0.150833
3)	deposit_type	0.147614
4)	arrival_date_month	0.072284
5)	total_of_special_requests	0.060666
6)	stays_in_week_nights	0.060448
7)	market_segment	0.055489
8)	previous_cancellations	0.053882
9)	stays_in_weekend_nights	0.037192
10)	customer_type	0.031932
11)	reserved_room_type	0.025205
12)	required_car_parking_spaces	0.021722
13)	adults	0.020920
14)	meal	0.018058
15)	hotel	0.013272
16)	distribution_channel	0.012901
17)	number_of_kids	0.009293
18)	previous_bookings_not_canceled	0.004957
19)	is_repeated_guest	0.002074

Conclusion

- Random Forest has the best average performance in all the metrics.
- ‘Lead_time’, ‘ADR’ are the most important features in both Xgboost and Random Forest.

  According to different models, the feature importance may change.
- Model Optimization(GridSearchCV) cannot improve the performance greatly all the time.

Yang Fei
Email: Sophia.fei0302@gmail.com
https://github.com/fysophia0302/SpringboardRepo/tree/master/Capstone_Project_2