Presentation to Hotel owners

Team Gold 3: Hai-Au Vo, Ritu Garg, Elaine Zhang, Charles Gay, & Naveed Safavi

BE BOUNDLESS

Agenda

- Project overview
- Data description
- Data modeling and methodologies
- Conclusion and business insights
- Reflection

Project overview

> **What:**

Explore what factors influence customer decision making about hotel reservation and answer three business questions:

1. How likely are customers to cancel their booking?
2. What factors influence the hotel type choice and how can each hotel type target to the right customers?
3. How likely are customers to reserve a parking space and how many spaces are needed if they do?

> **Who would benefit:**

Hotel business owners

> **How would the data help:**

With this data, hotel owners will know what customers are more likely to cancel and request parking space, and what market to target to gain higher revenue

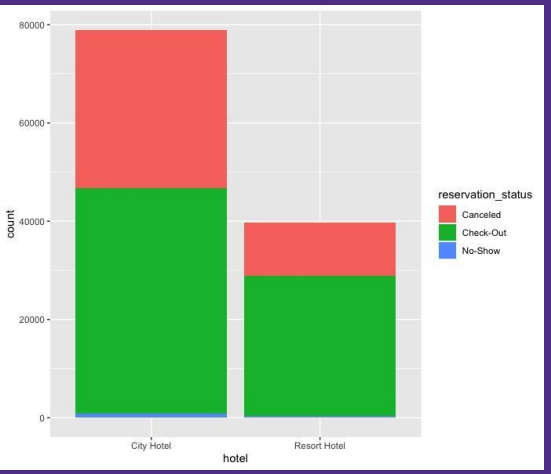
Data Description

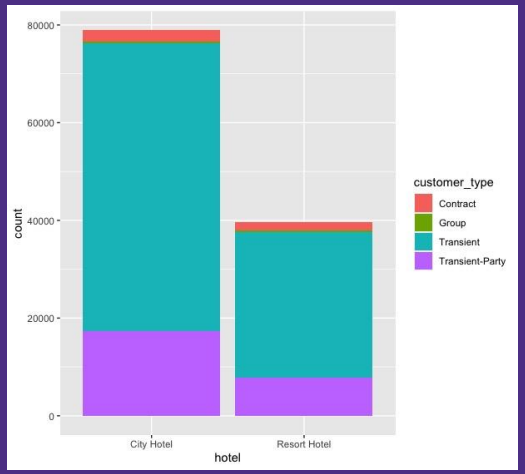
- Dataset covers 2 types of hotels
- Each observation represents a hotel booking
- Data coverage: July 2015 - August 2017, includes bookings that effectively arrived and canceled bookings
- 34 variables with various data types (num, int, factor)
- Personal information about customer (age, gender, nationality, etc.) are excluded

Data Description: Attributes

Variable	Type	Description
Hotel	Categorical	(H1 = Resort Hotel or H2 = City Hotel)
ADR	Numeric	Average Daily Rate -Calculated by dividing the sum of all lodging transactions by the total number of staying nights
Adults	Integer	Number of adults
Agent	Categorical	The ID of the travel agency that made the booking
ArrivalDateDayOfMonth	Integer	Day of the month of the arrival date
ArrivalDateMonth	Categorical	The month of arrival date with 12 categories: "January" to "December"
ArrivalDateWeek Number	Integer	Week number of the arrival date
ArrivalDateYear	Integer	Year of arrival date
AssignedRoomType	Categorical	Code for the type of room assigned to the booking. Sometimes the assigned room type differs from the reserved room type due to hotel operation reasons (e.g. overbooking) or by customer request. Code is presented instead of designation for anonymity reasons
Babies	Integer	Number of babies
BookingChanges	Integer	Number of changes/amendments made to the booking from the moment the booking was entered on the PMS until the moment of check-in or cancellation
Children	Integer	Number of children
Company	Categorical	The ID of the company/entity that made the booking or responsible for paying the booking. ID is presented instead of designation for anonymity reasons
Country	Categorical	Country of origin. Categories are represented in the ISO 3155–3:2013 format [6]
Customer type	Categorical	Type of booking, assuming one of four categories: Contract - when the booking has an allotment or other type of contract associated with it; Group – when the booking is associated with a group; Transient – when the booking is not part of a group or contract and is not associated with another transient booking; Transient-party – when the booking is transient but is associated with at least another transient booking
DaysInWaitingList	Integer	Number of days the booking was on the waiting list before it was confirmed to the customer
DepositType	Categorical	Indication on if the customer made a deposit to guarantee the booking. This variable can assume three categories: No Deposit – no deposit was made; In case no payments were found the value is "No Deposit". Non Refund - if the payment was equal or exceeded the total cost of stay Refundable – a deposit was made with a value under the total cost of the stay.
DistributionChannel	Categorical	Distribution channel. The term "TA" means "Travel Agents" and "TO" means "Tour Operators"
IsCanceled	Categorical	A value indicating if the booking was canceled (1) or not (0)
IsRepeatedGuest	Categorical	A value indicating if the booking name was a repeated guest (1) or not (0)
LeadTime	Integer	Number of days that elapsed between the entering date of the booking into the PMS and the arrival date
Market segment	Categorical	Market segment designation. In categories, the term "TA" means "Travel Agents" and "TO" means "Tour Operators"
Meal	Categorical	Type of meal booked. Undefined/SC – no meal package; BB – Bed & Breakfast; HB – Half board (breakfast and one other meal – usually dinner); FB – Full board (breakfast, lunch, and dinner)
PreviousBookingsNotCanceled	Integer	Number of previous bookings not canceled by the customer prior to the current booking
PreviousCancellations	Integer	Number of previous bookings that were canceled by the customer prior to the current booking
RequiredCardParkingSpaces	Integer	Number of car parking spaces required by the customer
ReservationStatus	Categorical	Reservation the last status, assuming one of three categories: Canceled – booking was canceled by the customer; Check-Out – customer has checked in but already departed; No-Show customer did not check-in and did not inform the hotel
ReservationStatusDate	Date	The date at which the last status was set. This variable can be used in conjunction with the *ReservationStatus* to understand when was the booking canceled or when did the customer checked-out of the hotel
ReservedRoomType	Categorical	Code of room type reserved. Code is presented instead of designation for anonymity reasons
StaysInWeekendNights	Integer	Number of weekend nights (Saturday or Sunday) the guest stayed or booked to stay at the hotel
StaysInWeekNights	Integer	Number of weeknights (Monday to Friday) the guest stayed or booked to stay at the hotel
TotalOfSpecialRequests	Integer	Number of special requests made by the customer

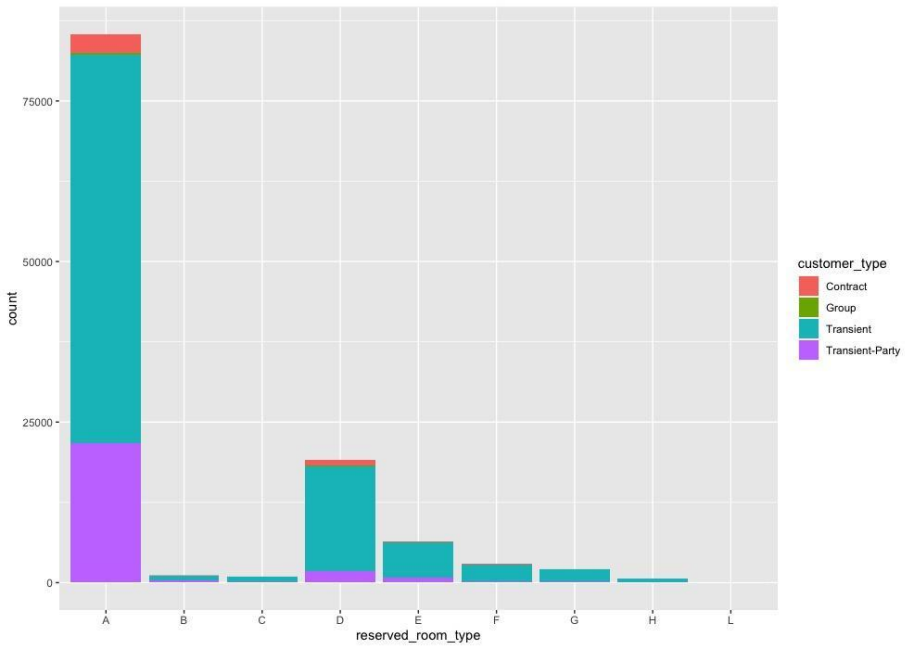
Exploratory data analysis





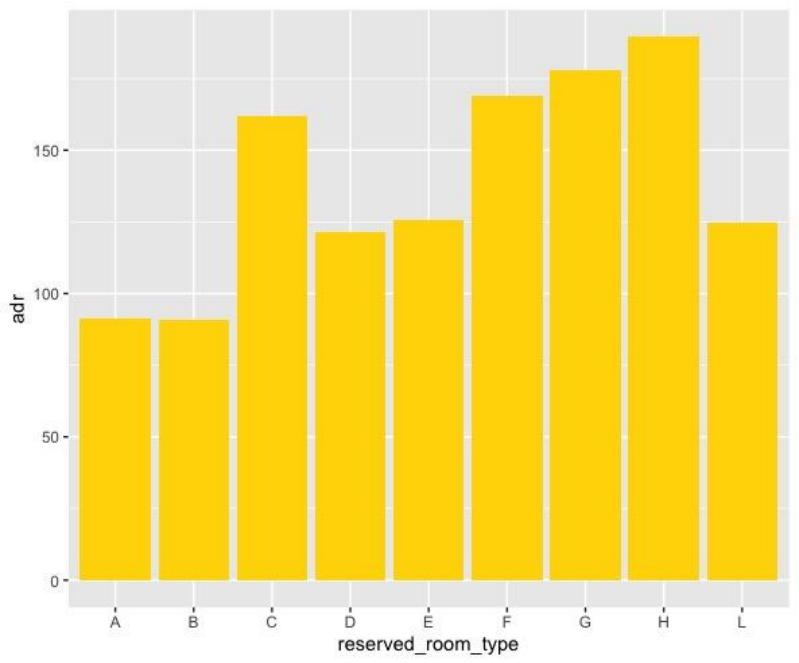
- City Hotel captures majority however, with higher cancelation rate proportionally
- transient customers make up the majority of customer type

Exploratory data analysis



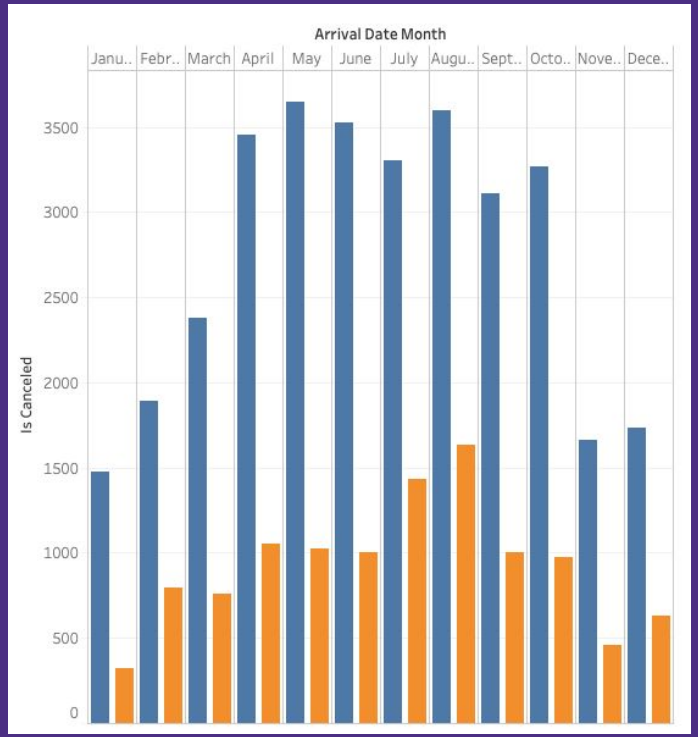
- City Hotel captures majority however, with higher cancelation rate proportionally
- transient customers make up the majority of customer type
- Popular room type: A, D, and E
- Least popular room type: H, C, B

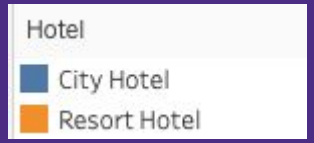
Exploratory data analysis



- City Hotel captures majority however, with higher cancelation rate proportionally
- transient customers make up the majority of customer type
- Popular room type: A, D, and E
- Least popular room type: H, C, B
- Highest ADR: H, D, F
- Lowest ADR: A,B,D

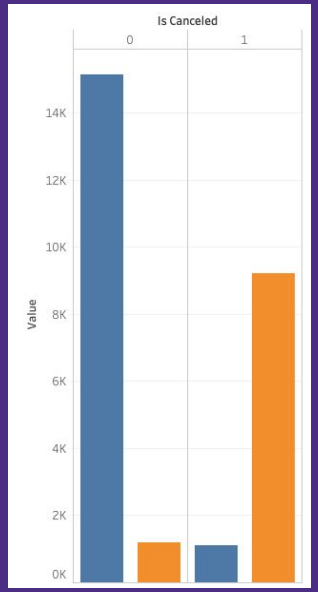
Exploratory data analysis

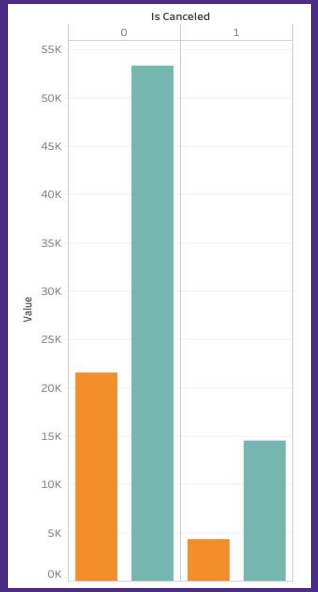


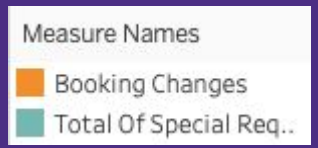


- cancellation rates lowest during winter months and highest during warmer months
- proportionally, resort hotels run a smaller rate of cancellations than city hotel

Exploratory data analysis



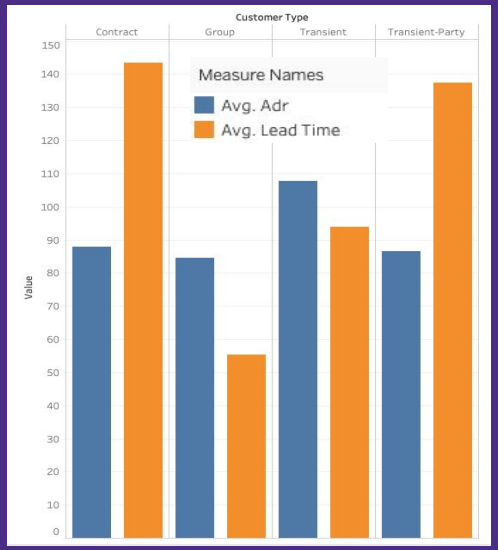


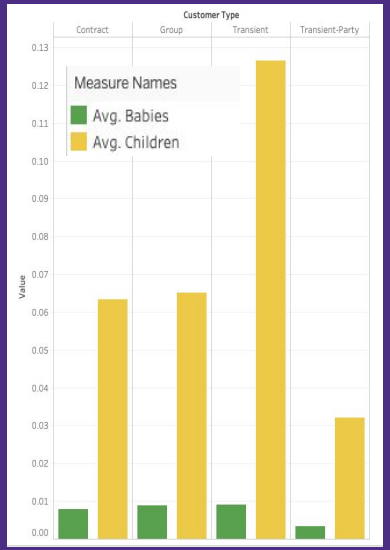


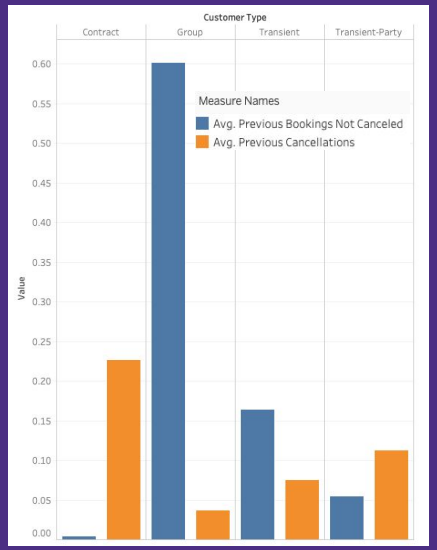
- Most likely to cancel:
  - history of frequent cancellations
  - Low numbers of special request
  - Low numbers of booking changes

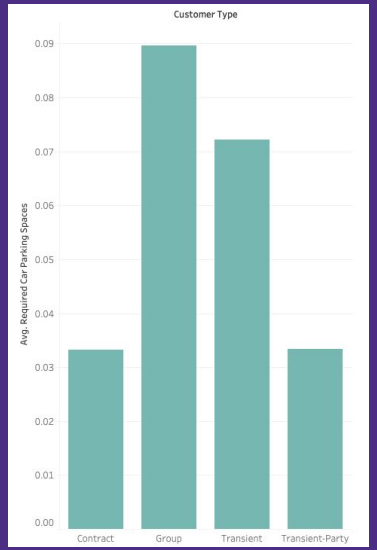
UNIVERSITY of WASHINGTON

Customer Profile









- **Contract:** average pay, tend to bring children + babies, highest rate of cancellations & lead time, less sparking space
- **Group:** lowest paying customers, ted to bring children + babies, cleanest record for cancellations, lowest lead time, require parking
- **Transient:** highest paying customer, more children + babies, tend to require parking space
- **Transient-party:** average pay, less likely to bring children + babies, tend to book further out, require less parking space

Question 1: Booking cancellation prediction

Methods used:

- Clustering
- Logistic regression
- Classification tree
- Random forest

Q1: Clustering

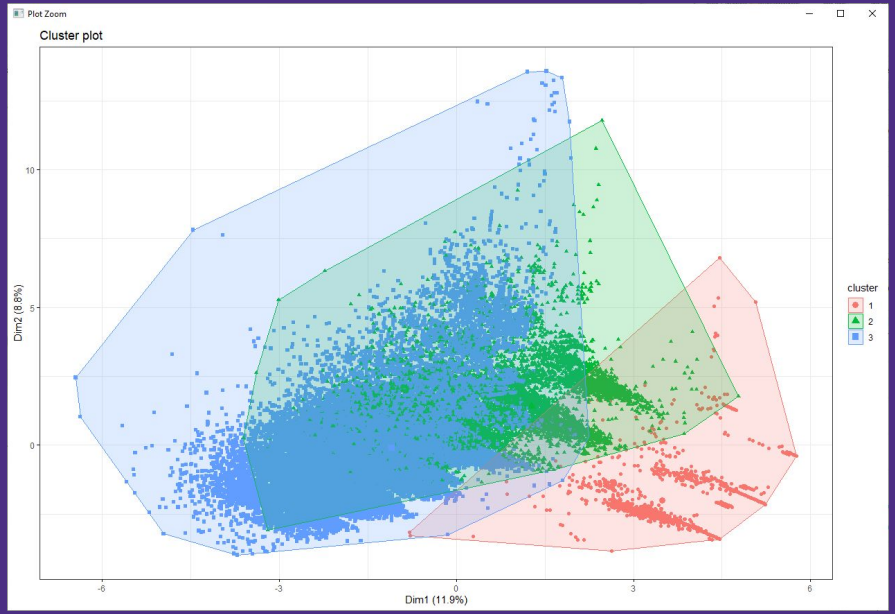
- This method gives us interesting insights about dataset attributes we might not have noticed and potentially reinforce conclusions
- Extra preprocessing steps to prepare the data for k-means (highly sensitive) - one hot encoding, normalizing
- Using the elbow method to determine the optimal number of clusters

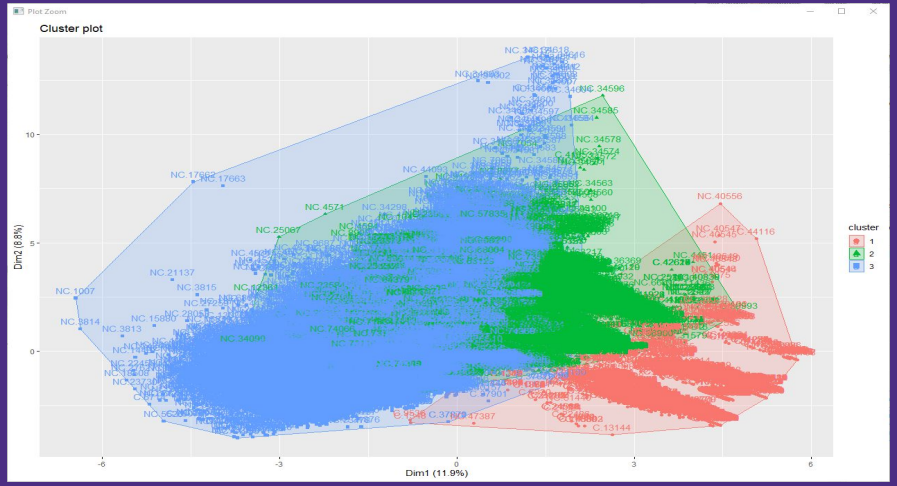
Cluster pre-processing

```
$ lead_time                      : int  7 13 14 14 0 9 85 75 23 35 ...
$ stays_in_weekend_nights        : int  0 0 0 0 0 0 0 0 0 0 ...
$ stays_in_week_nights           : int  1 1 2 2 2 2 3 3 4 4 ...
$ adults                         : int  1 1 2 2 2 2 2 2 2 2 ...
$ children                       : int  0 0 0 0 0 0 0 0 0 0 ...
$ babies                         : int  0 0 0 0 0 0 0 0 0 0 ...
$ is_repeated_guest              : int  0 0 0 0 0 0 0 0 0 0 ...
$ previous_cancellations         : int  0 0 0 0 0 0 0 0 0 0 ...
$ previous_bookings_not_canceled: int  0 0 0 0 0 0 0 0 0 0 ...
$ booking_changes                : int  0 0 0 0 0 0 0 0 0 0 ...
$ required_car_parking_spaces    : int  0 0 0 0 0 0 0 0 0 0 ...
$ total_of_special_requests      : int  0 0 1 1 0 1 1 0 0 0 ...
$ total_guests                   : int  1 1 2 2 2 2 2 2 2 2 ...
$ hotel_City Hotel               : int  0 0 0 0 0 0 0 0 0 0 ...
$ hotel_Resort Hotel             : int  1 1 1 1 1 1 1 1 1 1 ...
$ market_segment_Aviation        : int  0 0 0 0 0 0 0 0 0 0 ...
$ market_segment_Complementary   : int  0 0 0 0 0 0 0 0 0 0 ...
$ market_segment_Corporate       : int  0 1 0 0 0 0 0 0 0 0 ...
$ market_segment_Direct          : int  1 0 0 0 1 1 0 0 0 0 ...
$ market_segment_Groups          : int  0 0 0 0 0 0 0 0 0 0 ...
$ market_segment_Offline TA/TO   : int  0 0 0 0 0 0 0 1 0 0 ...
$ market_segment_Online TA       : int  0 0 1 1 0 0 1 0 1 1 ...
$ market_segment_Undefined       : int  0 0 0 0 0 0 0 0 0 0 ...
$ deposit_type_No Deposit        : int  1 1 1 1 1 1 1 1 1 1 ...
$ deposit_type_Non Refund        : int  0 0 0 0 0 0 0 0 0 0 ...
$ deposit_type_Refundable        : int  0 0 0 0 0 0 0 0 0 0 ...
$ customer_type_Contract         : int  0 0 0 0 0 0 0 0 0 0 ...
$ customer_type_Group            : int  0 0 0 0 0 0 0 0 0 0 ...
$ customer_type_Transient        : int  1 1 1 1 1 1 1 1 1 1 ...
$ customer_type_Transient-Party : int  0 0 0 0 0 0 0 0 0 0 ...
```

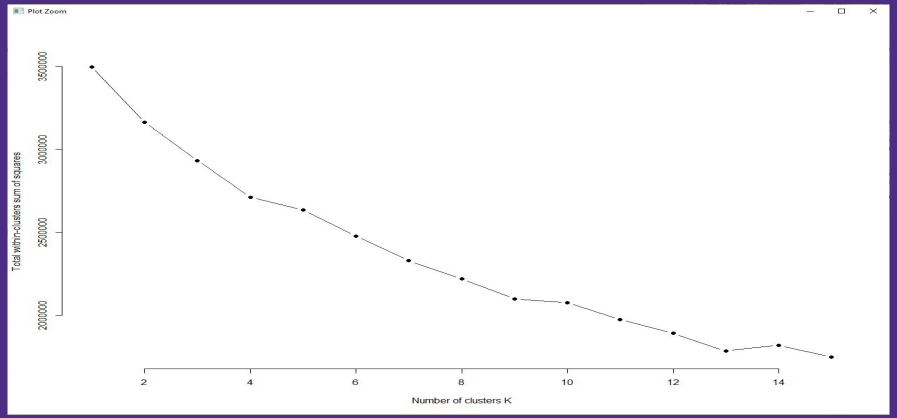
```
$ children                       : num  -0.262 -0.262 -0.262 -0.262 -0.262 ...
$ babies                         : num  -0.0873 -0.0873 -0.0873 -0.0873 -0.0873 ...
$ is_repeated_guest              : num  -0.174 -0.174 -0.174 -0.174 -0.174 ...
$ booking_changes                : num  -0.342 -0.342 -0.342 -0.342 -0.342 ...
$ required_car_parking_spaces    : num  -0.255 -0.255 -0.255 -0.255 -0.255 ...
$ total_of_special_requests      : num  -0.721 -0.721 0.54 0.54 -0.721 ...
$ total_guests                   : num  -1.3545 -1.3545 0.0391 0.0391 0.0391 ...
$ hotel_City Hotel               : num  -1.41 -1.41 -1.41 -1.41 -1.41 ...
$ hotel_Resort Hotel             : num  1.41 1.41 1.41 1.41 1.41 ...
$ market_segment_Aviation        : num  -0.0442 -0.0442 -0.0442 -0.0442 -0.0442 ...
$ market_segment_Complementary   : num  -0.0779 -0.0779 -0.0779 -0.0779 -0.0779 ...
$ market_segment_Corporate       : num  -0.215 4.655 -0.215 -0.215 -0.215 ...
$ market_segment_Direct          : num  2.92 -0.343 -0.343 -0.343 2.92 ...
$ market_segment_Groups          : num  -0.447 -0.447 -0.447 -0.447 -0.447 ...
$ market_segment_Offline TA/TO   : num  -0.504 -0.504 -0.504 -0.504 -0.504 ...
$ market_segment_Online TA       : num  -0.948 -0.948 1.055 1.055 -0.948 ...
$ market_segment_Undefined       : num  -0.00411 -0.00411 -0.00411 -0.00411 -0.00411 ...
$ deposit_type_No Deposit        : num  0.377 0.377 0.377 0.377 0.377 ...
$ deposit_type_Non Refund        : num  -0.375 -0.375 -0.375 -0.375 -0.375 ...
$ deposit_type_Refundable        : num  -0.037 -0.037 -0.037 -0.037 -0.037 ...
$ customer_type_Contract         : num  -0.188 -0.188 -0.188 -0.188 -0.188 ...
$ customer_type_Group            : num  -0.0694 -0.0694 -0.0694 -0.0694 -0.0694 ...
$ customer_type_Transient        : num  0.577 0.577 0.577 0.577 0.577 ...
```

Visualizations





	Cluster 1	Cluster 2	Cluster 3
Canceled	14493	5821	23842
Did not cancel (NC)	90	20840	53455



Interpreting results: Cluster centers

```
> km12$centers
   lead_time stays_in_weekend_nights stays_in_week_nights      adults   children      babies is_repeated_guest previous_cancellations
1  1.0144024              -0.31173677           -0.22269632 -0.08443436 -0.2599882 -0.08726022       -0.14882790              0.38237012
2  0.2657927              -0.00555500            0.01353531 -0.21240863 -0.1859743 -0.04824633       -0.02074887             -0.00316269
3 -0.2833937               0.06062131            0.03719252  0.08122355  0.1132500  0.03312196        0.03527221             -0.07102323
  previous_bookings_not_canceled booking_changes required_car_parking_spaces total_of_special_requests total_guests hotel_City Hotel
1                    -0.08432097    -0.323042800                 -0.25511109                -0.7188080   -0.2228270       0.45964591
2                    -0.05296374     0.196520453                 -0.12153527                -0.2700929   -0.2795342      -0.05370757
3                     0.03419544    -0.006782057                  0.09010309                 0.2288749    0.1320964      -0.06788112
  hotel_Resort Hotel market_segment_Aviation market_segment_Complementary market_segment_Corporate market_segment_Direct
1        -0.45964591              -0.04418334                  -0.07789249               -0.10395341            -0.3382694
2         0.05370757              -0.02802258                  -0.05899884                0.08857258            -0.1927601
3         0.06788112               0.01801041                   0.03506124               -0.01089055             0.1299968
  market_segment_Groups market_segment_Offline TA/TO market_segment_Online TA market_segment_Undefined deposit_type_No Deposit
1             1.2405307                     0.3491181               -0.9406122              -0.004107218              -2.6531108
2             0.5733871                     0.4003036               -0.6569183               0.014157354               0.3604336
3            -0.4317020                    -0.2040134                0.4041858              -0.004107218               0.3761289
  deposit_type_Non Refund deposit_type_Refundable customer_type_Contract customer_type_Group customer_type_Transient
1               2.6698843              -0.03698993             0.01706542        -0.06944322               0.3118841
2              -0.3742023               0.10719457             0.28647381         0.21874344              -1.7315105
3              -0.3745449              -0.02998566            -0.10198507        -0.06476348               0.5386332
  customer_type_Transient-Party
1                    -0.3268433
2                     1.6729749
3                    -0.5152465
```

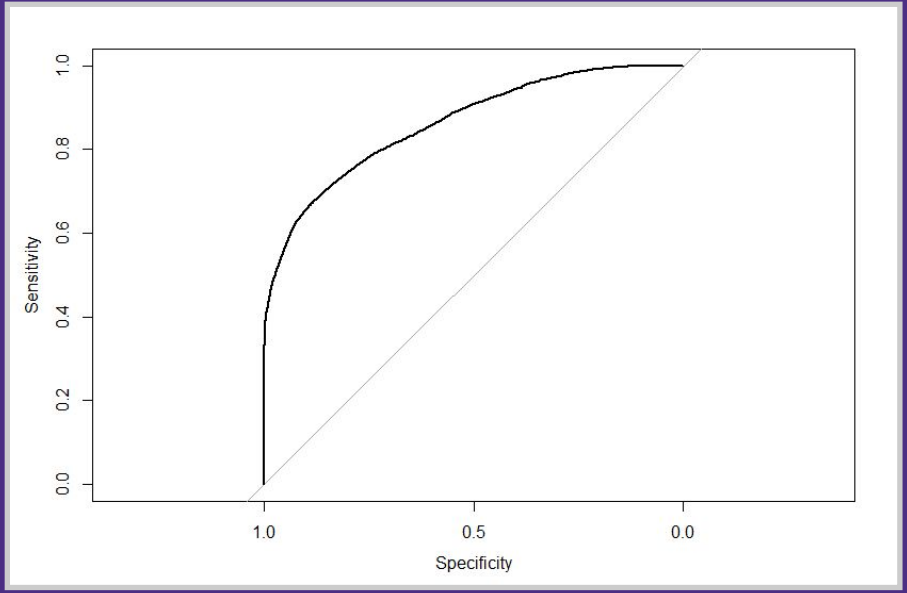
> Hotel type, lead time, previous cancellations are some interesting factors to note

Q1: Logistic regression

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904  -0.7444  -0.3574   0.2445   5.7443

Coefficients:
                                  Estimate   Std. Error z value             Pr(>|z|)
(Intercept)                     -1.5977704    0.0671077 -23.809 < 0.0000000000000002 ***
hotelResort Hotel                0.1746224    0.0231246   7.551 0.000000000000043072 ***
lead_time                        0.0035934    0.0001106  32.495 < 0.0000000000000002 ***
stays_in_weekend_nights          0.0374363    0.0107005   3.499             0.000468 ***
stays_in_week_nights             0.0385152    0.0056395   6.830 0.000000000008520582 ***
adults                           0.1513290    0.0202750   7.464 0.000000000000084053 ***
children                         0.2430730    0.0298475   8.144 0.000000000000000383 ***
babies                           0.3724246    0.1176959   3.164             0.001555 **
market_segmentAviation          -0.0859702    0.2102945  -0.409             0.682680
market_segmentComplementary      0.6012341    0.1899397   3.165             0.001549 **
market_segmentDirect             0.1571927    0.1229735   1.278             0.201156
market_segmentGroups             0.1659646    0.0946376   1.754             0.079484 .
market_segmentoffline TA/TO     -0.4662817    0.0979422  -4.761 0.000001928406692236 ***
market_segmentonline TA          0.8684587    0.0974283   8.914 < 0.0000000000000002 ***
market_segmentUndefined         -5.9515071 6627.1774623  -0.001             0.999283
distribution_channelDirect      -0.5537616    0.1112864  -4.976 0.000000649099229042 ***
distribution_channelGDS         -1.2887263    0.2644166  -4.874 0.000001094534111195 ***
distribution_channelTA/TO       -0.1238383    0.0843091  -1.469             0.141871
distribution_channelUndefined   21.2743275 6623.2532785   0.003             0.997437
is_repeated_guestY              -0.5849426    0.1050607  -5.568 0.000000025818004448 ***
previous_cancellations           2.7063716    0.0712264  37.997 < 0.0000000000000002 ***
previous_bookings_not_canceled  -0.4692777    0.0287424 -16.327 < 0.0000000000000002 ***
reserved_room_typeB              0.6003561    0.1214912   4.942 0.000000774992692637 ***
reserved_room_typeC              1.4302506    0.1577996   9.064 < 0.0000000000000002 ***
reserved_room_typeD              1.0934548    0.0545367  20.050 < 0.0000000000000002 ***
reserved_room_typeE              2.0205233    0.1132000  17.849 < 0.0000000000000002 ***
reserved_room_typeF              2.0780340    0.1676349  12.396 < 0.0000000000000002 ***
reserved_room_typeG              3.0202711    0.2543107  11.876 < 0.0000000000000002 ***
reserved_room_typeH              2.1675150    0.5235193   4.140 0.000034688643603647 ***
reserved_room_typeL            -10.1593823  177.6146086  -0.057             0.954387
assigned_room_typeB             -0.6534572    0.0966119  -6.764 0.000000000013447982 ***
assigned_room_typeC             -1.4930680    0.1205361 -12.387 < 0.0000000000000002 ***
assigned_room_typeD             -1.2765317    0.0520504 -24.525 < 0.0000000000000002 ***
assigned_room_typeE             -2.1025330    0.1095806 -19.187 < 0.0000000000000002 ***
assigned_room_typeF             -2.6476320    0.1587853 -16.674 < 0.0000000000000002 ***
assigned_room_typeG             -3.4743685    0.2463202 -14.105 < 0.0000000000000002 ***
assigned_room_typeH             -2.4350652    0.5106756  -4.768 0.000001857676500890 ***
assigned_room_typeI             -3.4081216    0.5404760  -6.306 0.000000000286750622 ***
assigned_room_typeK             -1.9990078    0.3951800  -5.058 0.000004226243223339 ***
assigned_room_typeL             22.4074286  370.1424343   0.061             0.951728
booking_changes                 -0.3532588    0.0183510 -19.250 < 0.0000000000000002 ***
deposit_typeNon Refund           5.4568074    0.1356275  40.234 < 0.0000000000000002 ***
deposit_typeRefundable          -0.0515132    0.2642851  -0.195             0.845459
days_in_waiting_list            -0.0003464    0.0005937  -0.584             0.559532
customer_typeContract           -0.8709972    0.0635262 -13.711 < 0.0000000000000002 ***
customer_typeGroup              -1.1939440    0.2097060  -5.693 0.000000012452018701 ***
customer_typeTransient-Party    -0.4935477    0.0334230 -14.767 < 0.0000000000000002 ***
adr                              0.0035930    0.0002436  14.750 < 0.0000000000000002 ***
required_car_parking_spaces    -38.9657553   81.3231969  -0.479             0.631834
total_of_special_requests       -0.7254480    0.0138512 -52.374 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 20727  4999
         1  1570  8272

               Accuracy : 0.8153
                 95% CI : (0.8112, 0.8193)
    No Information Rate : 0.6269
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.5834

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.6233
            Specificity : 0.9296
         Pos Pred Value : 0.8405
         Neg Pred Value : 0.8057
             Prevalence : 0.3731
         Detection Rate : 0.2326
   Detection Prevalence : 0.2767
      Balanced Accuracy : 0.7765

       'Positive' Class : 1
```

Q1: Logistic regression



	threshold	specificity	sensitivity
1	0.1	0.2879038	0.9768747
2	0.2	0.5034631	0.9059064
3	0.3	0.7003888	0.8074153
4	0.4	0.8399392	0.7128668
5	0.5	0.9230975	0.6259004

Q1: Logistic regression

```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 20727  4999
         1  1570  8272

               Accuracy : 0.8153
                 95% CI : (0.8112, 0.8193)
    No Information Rate : 0.6269
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.5834

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.6233
            Specificity : 0.9296
         Pos Pred Value : 0.8405
         Neg Pred Value : 0.8057
             Prevalence : 0.3731
         Detection Rate : 0.2326
   Detection Prevalence : 0.2767
      Balanced Accuracy : 0.7765

       'Positive' Class : 1
```

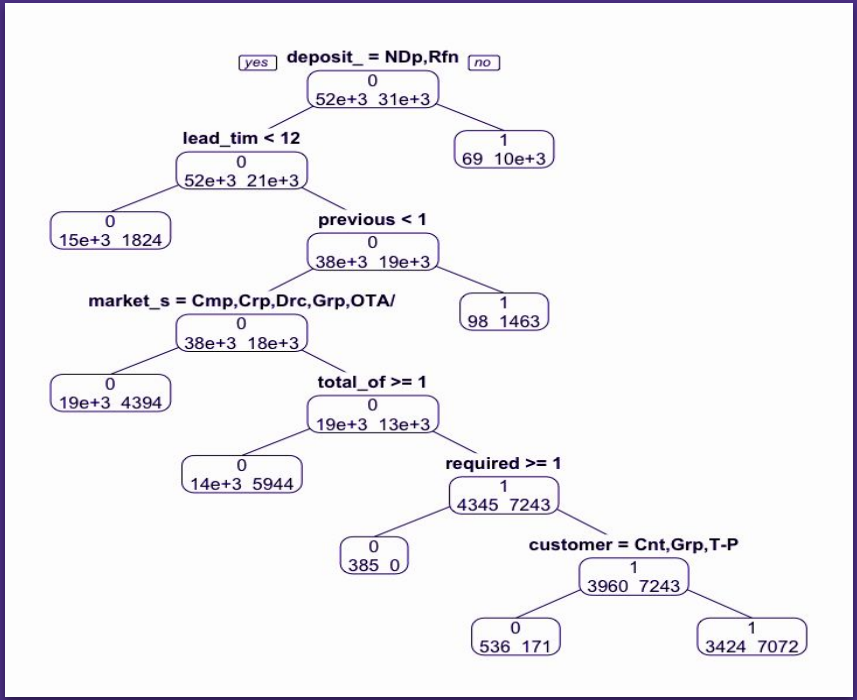
```
Confusion Matrix and Statistics

          Reference
Prediction     0     1
         0 19730  4265
         1  2567  9006

               Accuracy : 0.8079
                 95% CI : (0.8038, 0.812)
    No Information Rate : 0.6269
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.5785

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.6786
            Specificity : 0.8849
         Pos Pred Value : 0.7782
         Neg Pred Value : 0.8223
             Prevalence : 0.3731
         Detection Rate : 0.2532
   Detection Prevalence : 0.3254
      Balanced Accuracy : 0.7817

       'Positive' Class : 1
```

Q1: Classification tree

Correlation Matrix for important predictors

Attribute	Correlation Value
deposit_type	0.468675519
lead_time	0.292875656
previous_cancellations	0.110139263
market_segment	0.059419331
booking_changes	-0.144831563
required_car_parking_spaces	-0.195701443
total_of_special_requests	-0.234877003

Q1: Classification Tree



Cancelled:

- deposit_type= Non Refundable
- deposit_type= No Deposit OR Refundable AND Lead_time>=12 and previous_cancellations>=1

Not Cancelled:

- deposit_type= No Deposit or Refundable AND Lead_time<12

Q1: Classification tree

Confusion matrix:

		Actual	
		0	1
Predicted	0	21102	5138
	1	1482	8041

'Positive' Class : 1

Insights

- Customers who either don't pay any deposit or pay only refundable deposit are more likely to cancel their bookings in cases where it was made more than 11 days prior to the arrival date with some cancellation history.
- Customers who either don't pay any deposit or pay only a refundable deposit and made bookings within 11 days from the arrival date are likely to check-in

Future Scope

Pruning the tree by controlling some of its parameters such as the minimum number of observations that must exist in a node in order for a split to be attempted or the minimum number of observations in any terminal leaf .This can help take away enough levels of complexity and maximize accuracy in predicting the cancellations.

Q1: Random Forest

Confusion matrix:

		Actual	
		0	1
Predicted	0	21028	3646
	1	1556	9533

'Positive' Class : 1

Insights

- Random Forest classification model improves test data prediction accuracy by 3% to 85.5%.
- The sensitivity of favorable class i.e Hotel Booking cancellations improves by 11% when compared with classification tree.

Future Scope

Hyperparameter tuning by changing number of variables tried at each split and increasing number of trees

Model Comparison

- Random forest gave most accurate results with an accuracy of 85.5%
- It also gave the highest sensitivity of 0.72

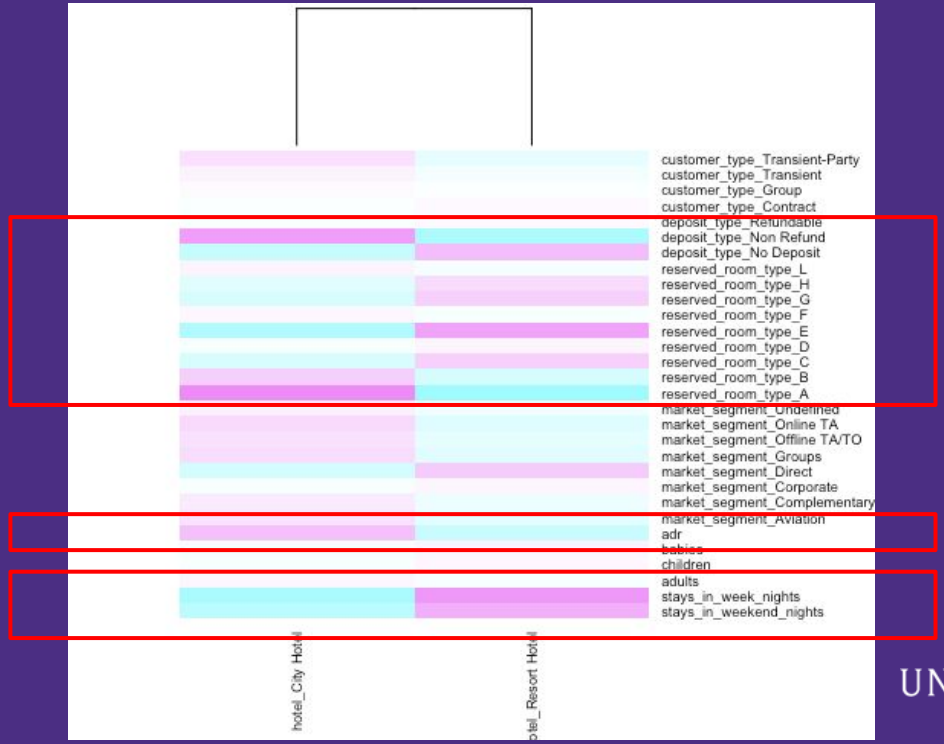
	Logistic Regression	Decision Tree	Random Forest
Accuracy	80.5%	81.5%	85.5%
Sensitivity	0.68	.61	.72
Specificity	0.88	.93	.93

Question 2:
Hotel-type choice &Target Market

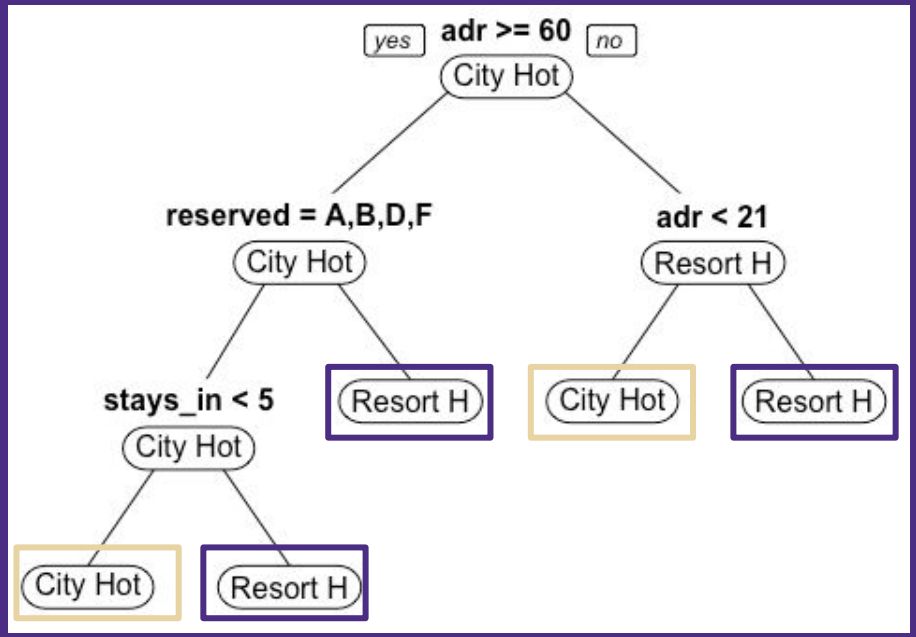
Methods used:

> Correlation Matrix and Decision Tree

Q2: Correlation Matrix



Q2: Decision Tree



City hotel:

1. adr < 21
2. adr >=60 & reserved = A,B,D,F, & and stays_in < 5

Resort hotel:

1. 21<adr < 60
2. adr >= 60 & reserved = C,E,G,H, L
3. adr >=60 & reserved = A,B,D,F & stays_in >=5

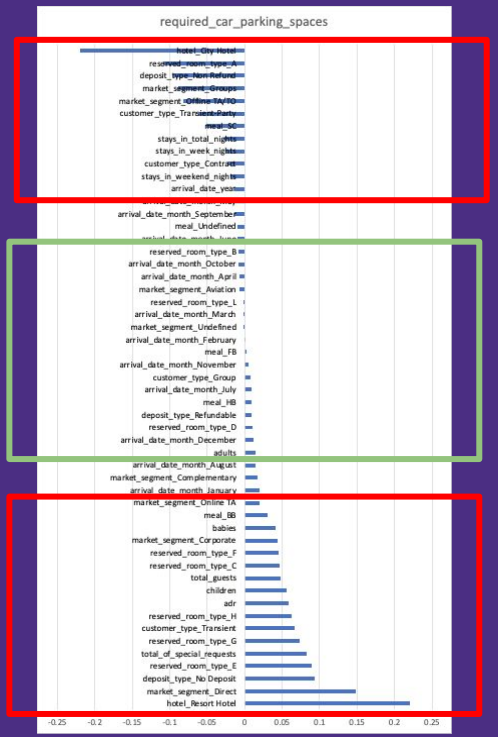
Question 3: Parking Space Request

Methods used:

> Correlation Matrix and Multiple Linear Regression

Q3: Correlation Matrix

arrival_date_month
reserved_room_type_B
market_segment_Aviation
reserved_room_type_L
market_segment_Undefined
meal_FB
customer_type_Group
meal_HB
deposit_type_Refundable
reserved_room_type_D

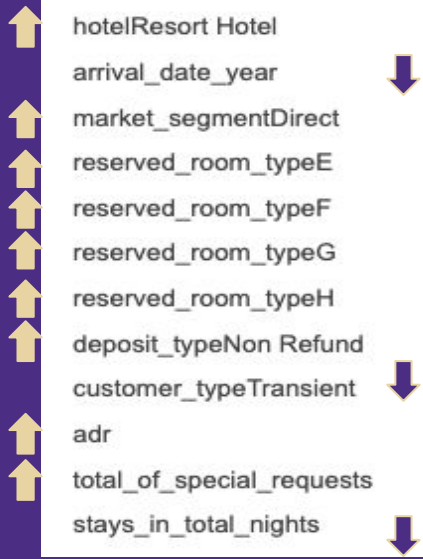


hotel_City Hotel
reserved_room_type_A
deposit_type_Non Refund
market_segment_Groups
market_segment_Offline TA/TO
customer_type_Transient-Party
meal_SC

children
adr
reserved_room_type_H
customer_type_Transient
reserved_room_type_G
total_of_special_requests
reserved_room_type_E
deposit_type_No Deposit
market_segment_Direct
hotel_Resort Hotel

Q3: Multiple Linear Regression

(results snapshot)



	Estimate	Std. Error	t value		Pr(>|t|)	
(Intercept)	27.92303548	2.63283901	10.606	<	2E-16	***
hotelResort Hotel	0.10705274	0.0021195	50.509	<	2E-16	***
arrival_date_year	-0.01384485	0.00130624	-10.599	<	2E-16	***
market_segmentDirect	0.0744188	0.02032448	3.662		0.000251	***
reserved_room_typeE	0.03781168	0.00413362	9.147	<	2E-16	***
reserved_room_typeF	0.02807106	0.00659717	4.255		2.09303E-05	***
reserved_room_typeG	0.06008456	0.00785978	7.645		2.12E-14	***
reserved_room_typeH	0.12353724	0.01337679	9.235	<	2E-16	***
deposit_typeNon Refund	-0.01730363	0.00366768	-4.718		2.3878E-06	***
customer_typeTransient	0.01983942	0.00506856	3.914		9.07841E-05	***
adr	0.00023103	0.00002329	9.918	<	2E-16	***
total_of_special_requests	0.0167103	0.00123637	13.516	<	2E-16	***
stays_in_total_nights	-0.00863455	0.00037138	-23.25	<	2E-16	***

Business Insights

Business Questions	**Booking cancellation decision**	**Hotel type target market**	**Parking space request prediction**
Data Analysis Model	Clustering Logistic regression Classification tree Random forest	Correlation analysis Decision tree	Correlation analysis Multiple linear regression
Key Insights	Customers who either **don't pay any deposit** or pay only **refundable deposit,** booked **11 days prior to the arrival date** and with some **cancellation history** are more likely to cancel	1) City hotel should target at customers with **lower budget, cheaper room type, less staying nights.** 2) Resort hotel should target to customers with **higher budget, higher end room type or staying longer in cheaper rooms**.	1) **Resort hotels** should expect more people to request parking spaces, and more space needed. 2) Hotels in general should expect a higher need from customers **travelling with babies** and have a **reserved room type of H**.
Model Evaluation Metrics	Highest Accuracy: 85.5%, Highest Sensitivity: 0.72	Accuracy: 82.6%	RMSE = 0.2372
Future Improvement	Need customer reviews data to understand the reason behind cancellations.	Create different customer persona profile based on hotel choosed.	Need more data to understand the real needs for parking: holiday season, parking fee...

Reflection

Key challenges:

- Converting large categorical variables for k-means
- Using common preprocessing steps
- Identifying and eliminating outliers

Potential Improvements:

- Acquire even more data

Thank You
Questions?

UNIVERSITY *of* WASHINGTON