Airbnb Data Challenge

Yifei Guo

Data Exploration and Cleaning

Understand Data - dataset summarization

- Contacts:
  - After cleaning the contacts dataset, there are a total of 27,856 observations with 14 variables.
    - All inquiries are made between 01/01/2016 and 06/30/2016
    - The dataset have 22,545 unique guests, 8,956 unique hosts, and 12,812 unique listings.
    - Most inquiries are made through Contact me channel
    - The majority guests are new user
    - The average number of guests per inquiry is 2.78 guests
- Listings:
  - There are a total of 12,997 listings with 4 variables.
    - There are 3 distinct room types and 68 distinct neighborhoods included
    - The average total review per listing is 7
- Users:
  - There are a total of 31,457 unique users with 3 variables.
    - The guests are from 121 different countries, and the vast majority of guests are from Brazil
    - The hosts are from 26 different countries
    - The average words in user profile is 16

Understand Data - business process

Assumptions:

1. For inquiries through ‘Contact Me’ Channel:
    - The first interaction time <= reply time <= accepted time <= booking time
2. For inquiries through ‘Book It’ Channel:
    - The first interaction time <= reply time <= accepted time = booking time
    - The booking time cannot be null if the inquiry was accepted
3. For inquiries through ‘Instant Booking’ Channel:
    - The first interaction time = reply time = accepted time = booking time
    - If an inquiry was sent, then its automatically accepted and booked
4. For all contact channels:
    - The inquiry time should not be later than the check in date
    - The check in date should not be later than the check out date
5. No negative value for total_reviews, m_interactions,m_first_message_length_in_characters, and words_in_user_profile
6. The m_guests cannot be less than one
7. No duplicate records in each table

Understand Data - data sanity checking rules

- ‘Contact Me’ Channel:

  The first interaction time <= reply time <= accepted time <= booking time

  - **Observation**:
    - There are 279 inquiries(2.17%) with reply time earlier than the first interaction time
      - The time lag between reply time and the first interaction time is between 1 second and 70 seconds, the mean time lag is 51.36 seconds
    - Out of the 279 inquiries, 209(1.63%) of them with accepted time earlier than the first interaction time
      - The time lag between accepted time and the first interaction time is between 1 second and 70 seconds, the mean time lag is 59 seconds
  - **Assumptions**:
    - Since the time lag is around one minute, it might be due to system delay
  - **Treatment**:
    - Populate the reply time and accepted time with the first interaction time for these records

Understand Data - data sanity checking rules

- ‘Book It’ Channel:

  The first interaction time <= reply time <= accepted time = booking time

  - **Observation**:
    - There are 257 inquiries(3.07%) with accept time but no booking time
    - There are 173 inquiries(2.07%) with reply time earlier than the first interaction time
      - The time lag between reply time and the first interaction time is between 1 second and 70 seconds, the mean time lag is 35.70 seconds
    - Out of the 173 inquiries, 122 (1.46%)of them with accepted time earlier than the first interaction time
      - The time lag between accepted time and the first interaction time is between 1 second and 70 seconds, the mean time lag is 40.80 seconds
  - **Assumptions**:
    - System Delay
  - **Treatment**:
    - Populate the booking time with accepted time for records without booking time and booking time not equal accepted time
    - Populate the reply time and accepted time with the first interaction time for these records

Understand Data - data sanity checking rules

- ‘Instant Booking’ Channel:

  The first interaction time = reply time = accepted time = booking time

  - **Observation**:
    - There are 3192 inquiries(47.70%) with inconsistent timestamps
    - Out of the 3192 inquiries, 3104 (46.38%) of them with the reply time not equal to the first interaction time
    - Out of the 3192 inquiries, 195 (2.91%) of them with the accepted time not equal to the reply time
    - Out of the 3192 inquiries, 14 (0.21%) of them with the booking time not equal to the accepted time
  - **Assumptions**:
    - System Delay
  - **Treatment**:
    - Populate all the timestamps with the first interaction time for these records

Understand Data - data sanity checking rules

- All Contact Channel:

  The inquiry time should not be later than the check in date

  The check in date should not be later than the check out date

  - **Observation:**
    - There are 26 inquiries(0.09%) with inquiry time later than the check in date
    - There are 0 inquiries(0%) with check in date later than the check out date
  - **Assumptions:**
    - System Errors
  - **Treatment:**
    - Filter out these records from analysis

Understand Data - data sanity checking rules

- **No negative value for total_reviews, m_interactions,m_first_message_length_in_characters, and words_in_user_profile**
  - Observation:
    - There are 41 negative total_reviews in the listings dataset
  - Treatment:
    - Filter out these records from analysis
- **m_guests cannot be less than one**
  - Observation:
    - There are 4 zero m_guest in the contacts dataset
  - Treatment:
    - Filter out these records from analysis
- **Remove duplicates in three datasets**
  - Observation:
    - There are 68 duplicates in the Users dataset.
  - Treatment:
    - We kept the first occurrence value as unique and drop the rest of the same values as duplicate.

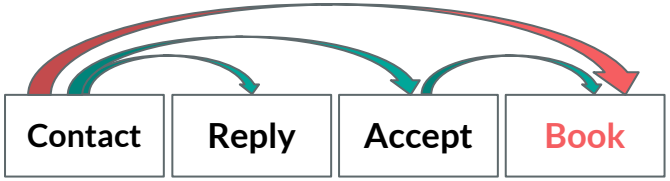
Understand Data - data sanity checking results

Before/After	inquiries	replied	accepted	booked
Contact Me	12,828/12,809	11,659/11,642	5,482/5,474	911/905
Book It	8,366/8,360	7,503/7,499	4,240/4,237	3,983/4,237
Instant Book	6,693/6,687	6,693/6,687	6,693/6,687	6,693/6,687
Total	27,887/27,856	25,855/25,828	16,415/16,398	11,587/11,829

Before/After	Records
Listings	13,038/12,997
Users	31,525/31,457

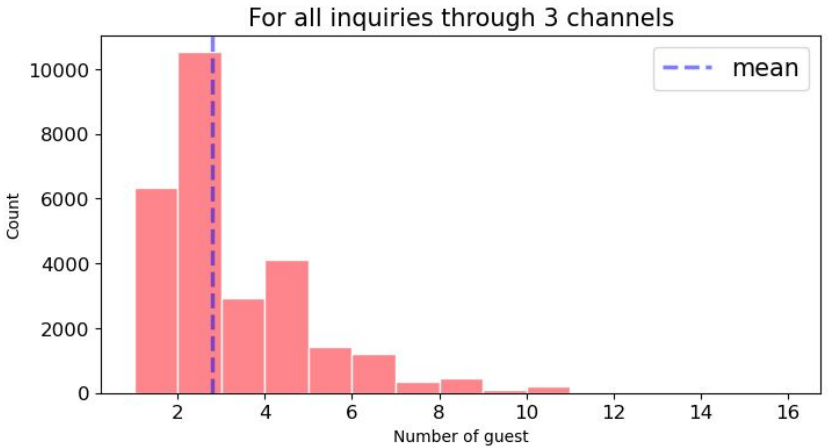
Understand Data - key metrics summarization

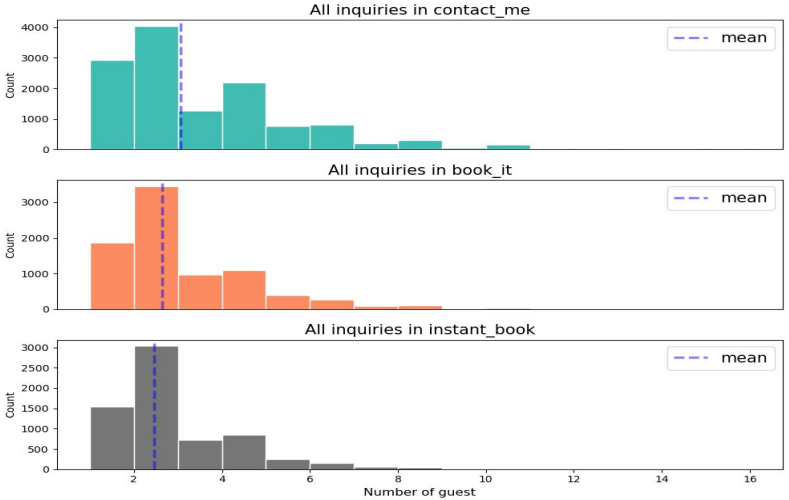
Metric	Conversion Rate
Contact to Reply Ratio	0.927
Contact to Accept Ratio	0.589
Contact to Book Ratio	**0.425**
Accept to Book Ratio	0.721



Exploratory Data Analysis

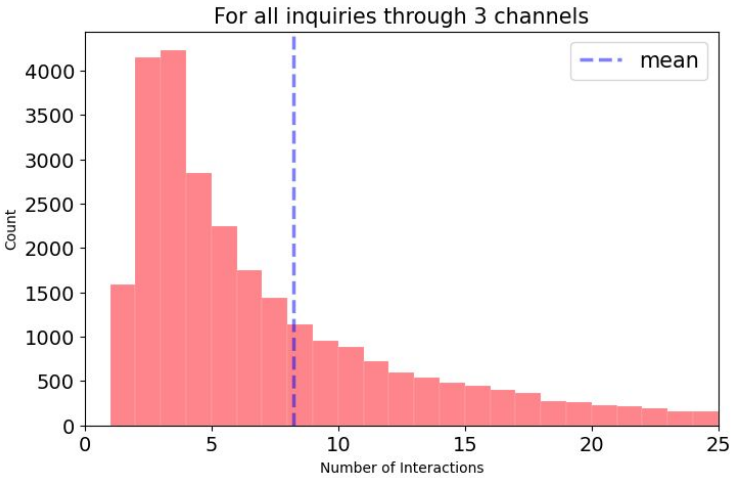
EDA - Group Size

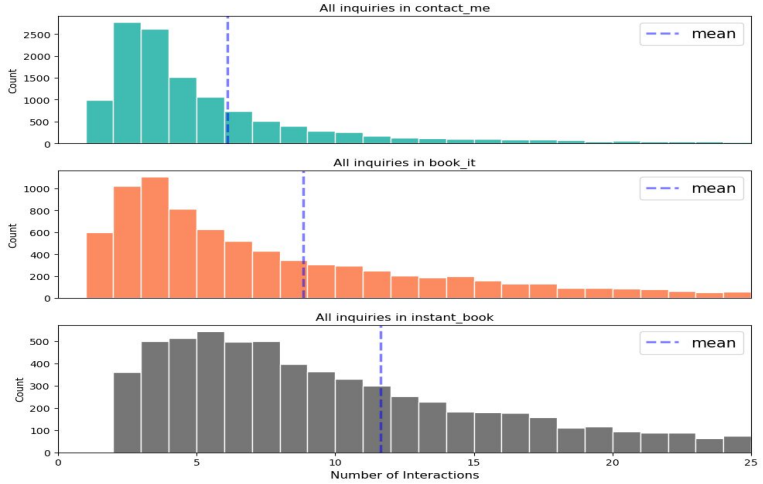




- The group size ranges from 1 to 16 people; about 95% of the inquiries were made for group sizes of six or less
- Most inquiries were made for a group of 2 people, which is about 37.9% of all inquiries
- As the group size increase, more inquiries were made through ‘contact me’ channel

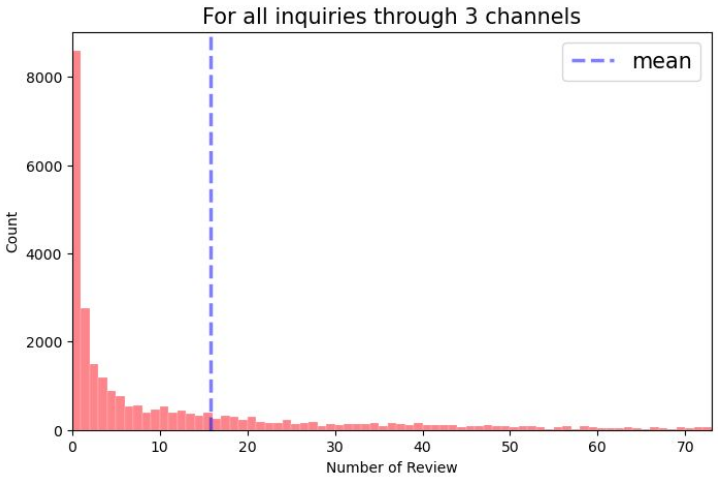
EDA - Number of Interactions

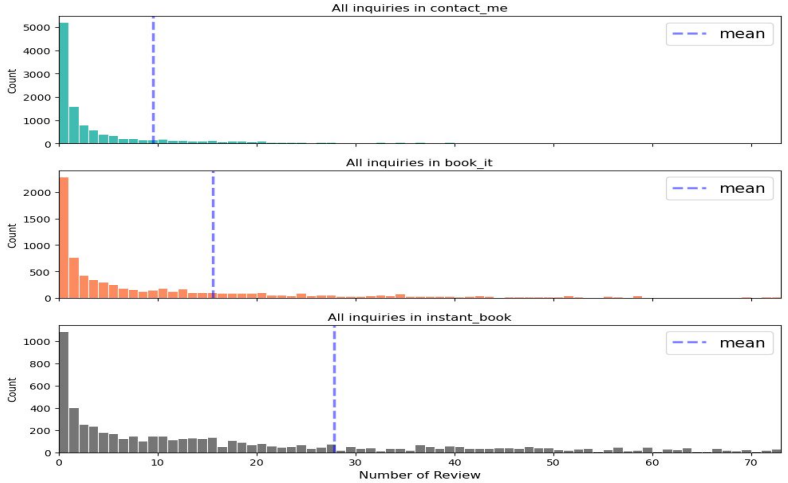




- The number of interactions ranges from 1 to 410 times, about 95% of the inquiries have 25 or less interactions
- The number of interactions peaks at 3, which is about 15.2% of all inquiries
- The average number of interactions for inquiries made by ‘contact me’ channel is 6, 8 for ‘book_it’ and 11 for ‘instant booking’ channel

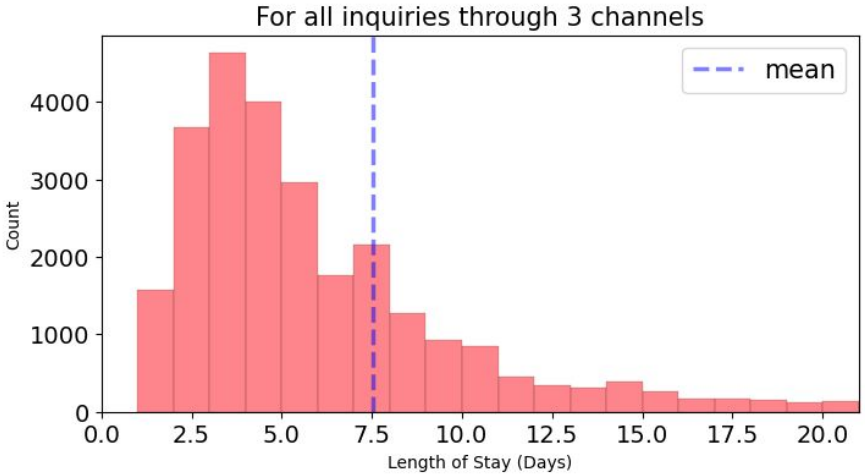
EDA - Number of Review

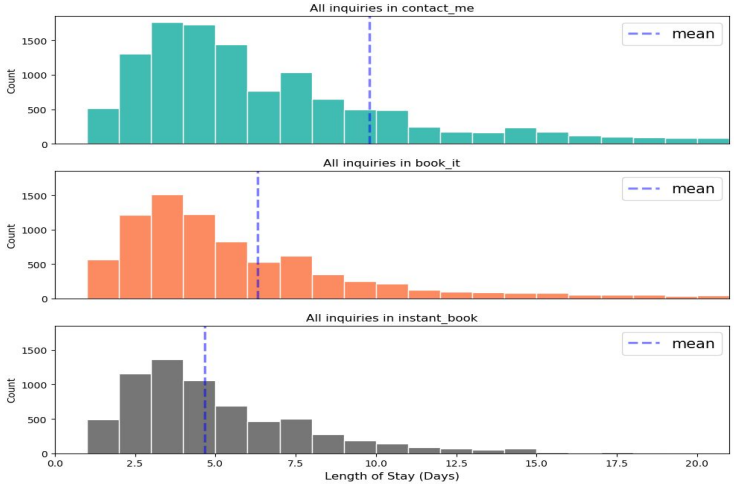




- The number of total reviews ranges from 1 to 268, about 50% of the inquiries have 3 or less reviews
- The average number of review is 16 for all listings
- Listings booked use ‘instant book’ have the highest average total reviews, followed by ‘book it’ and ‘contact me’ channel

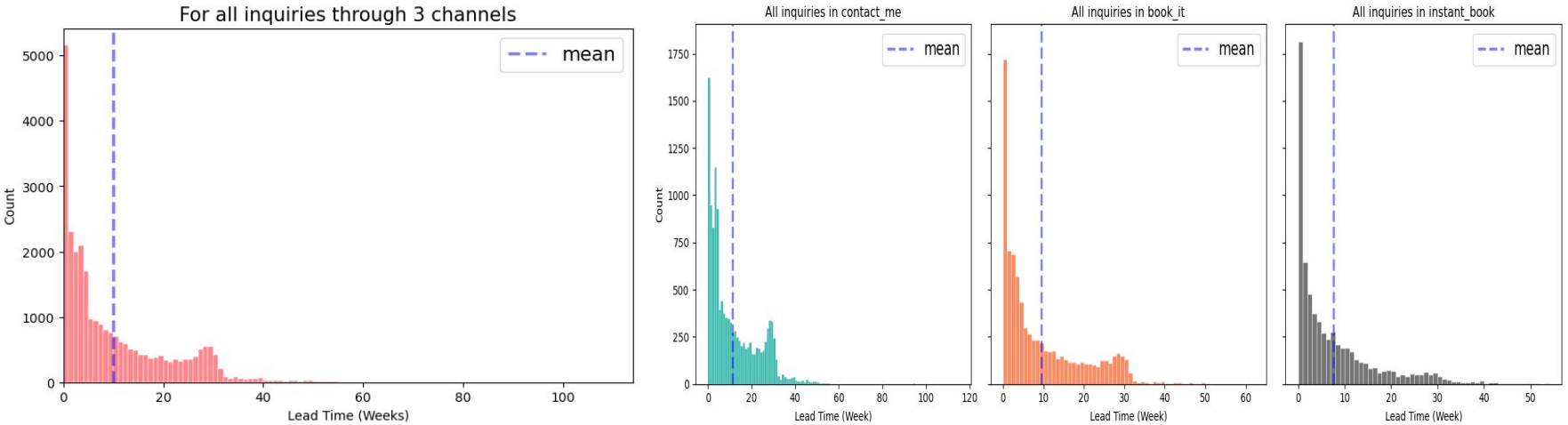
EDA - Length of Stay





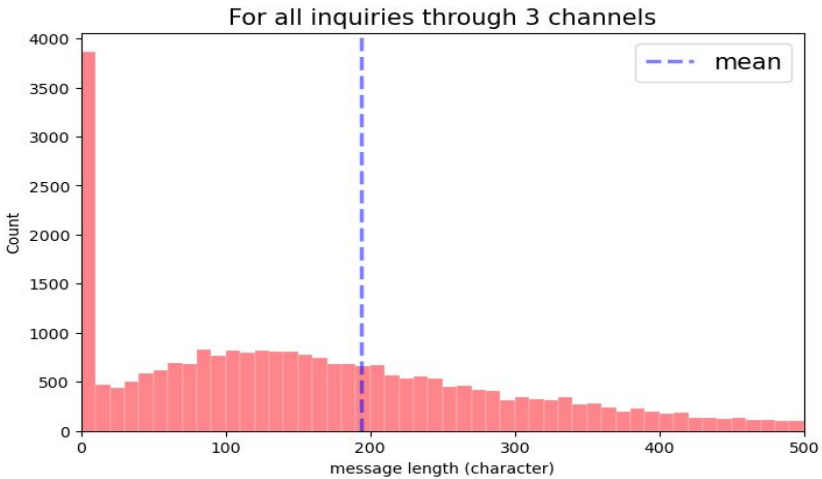
- The length of stay ranges from 1 day to 365 days; about 95% of the inquiries have a length of stay <=21 days
- Around 16.7% of the inquiries were made for a three-day trip, which is the most frequent length of stay in the dataset
- As the length of stay increase, fewer inquiries were made through ‘instant book’ channel and ‘book it’ channel; more through ‘contact_me’ channel

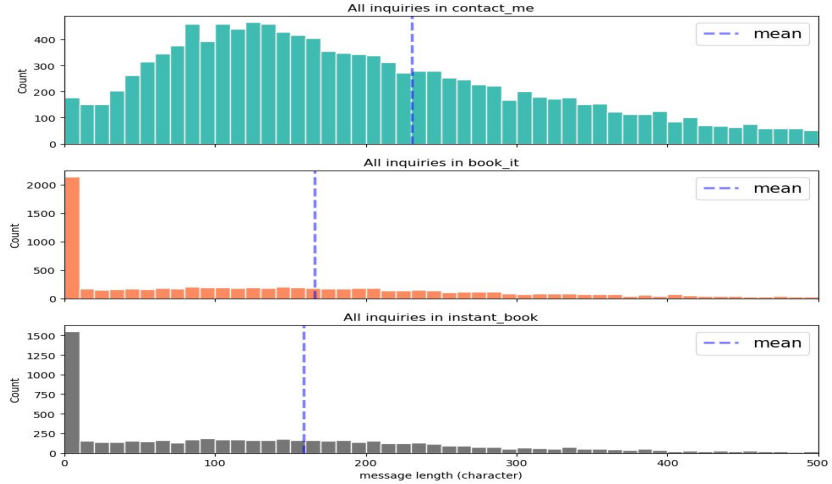
EDA - Booking Lead Time



- The booking lead time ranges from 0 to 114 weeks, about 95% of inquires have lead time <=29 weeks
- The average lead time is 9 weeks across all channels
- About 4.2% of all inquiries are made one day before the check-in date
- The average lead time in the ‘instant book’ channel is the smallest among three channels

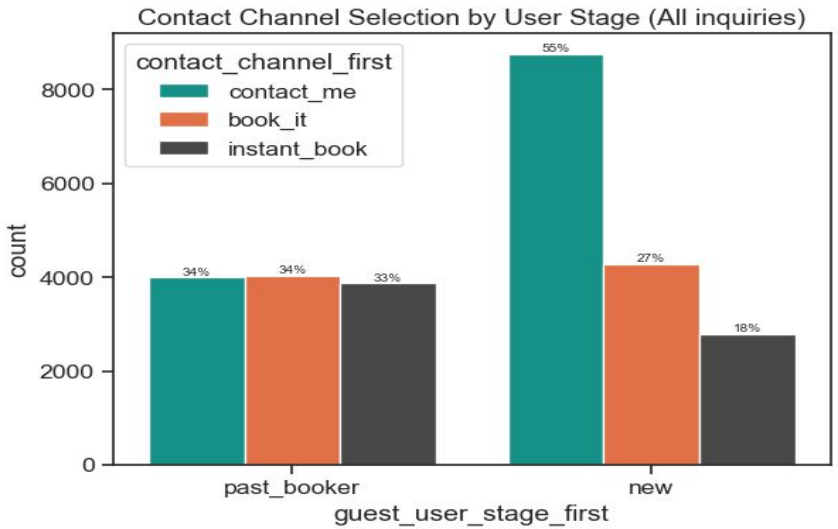
EDA - First message length (character)





- The message length ranges from 0 to 2341 characters, about 95% of inquires have message <=539 characters
- The average message length across all inquiries are 193 words
- 'book it' and 'instant book' channel's message length peak at zero character, and their average message length are lower than that from 'contact me' channel

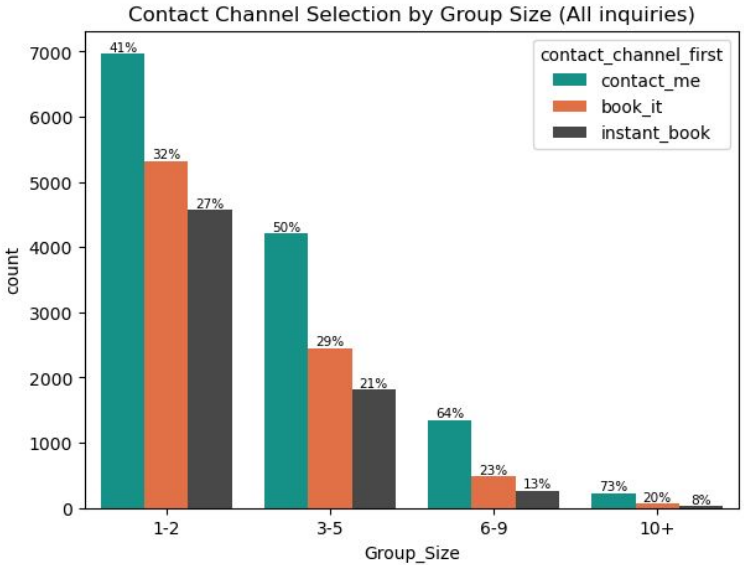
EDA - channel selection by user stage & success rate

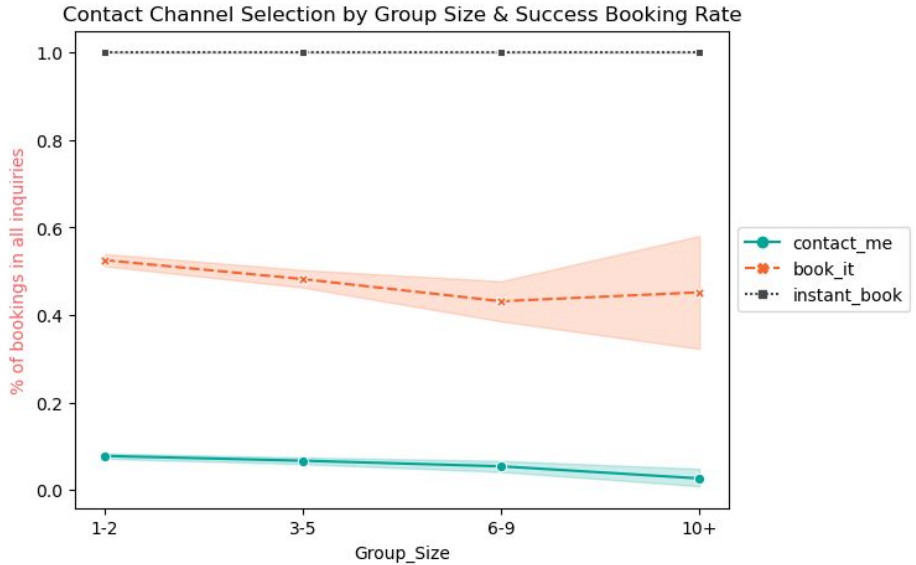


Contact Channel	User Stage	Reply Rate(%)	Accept Rate(%)	Booking Rate(%)
Contact me	New	91%	42%	**6%**
	Past Booker	90%	43%	**9%**
Book it	New	89%	49%	**49%**
	Past Booker	90%	53%	**53%**
Instant book	New	100%	100%	**100%**
	Past Booker	100%	100%	**100%**

- New users are more likely to use 'contact me' channel, 55% of new users inquiries are made through the 'Contact_me' channel compared to only 18% choose the 'instant book' channel
- There is no visible channel preference for past bookers
- The booking rate for past bookers are higher than new users in both the 'contact_me' channel and the 'book_it' channel

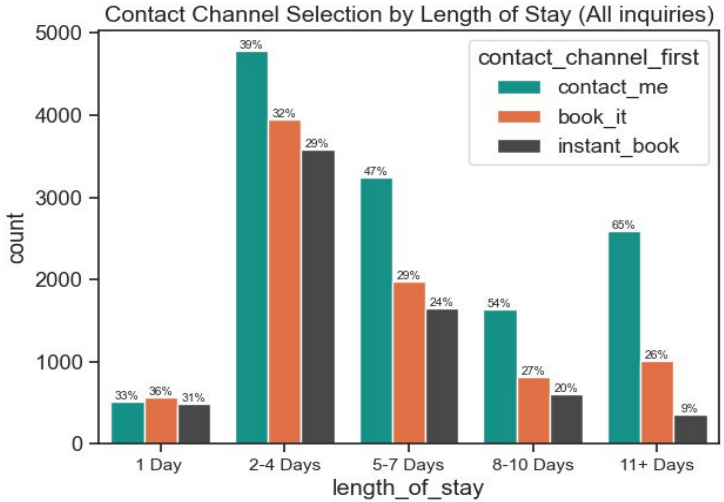
EDA - channel selection by group size & success rate

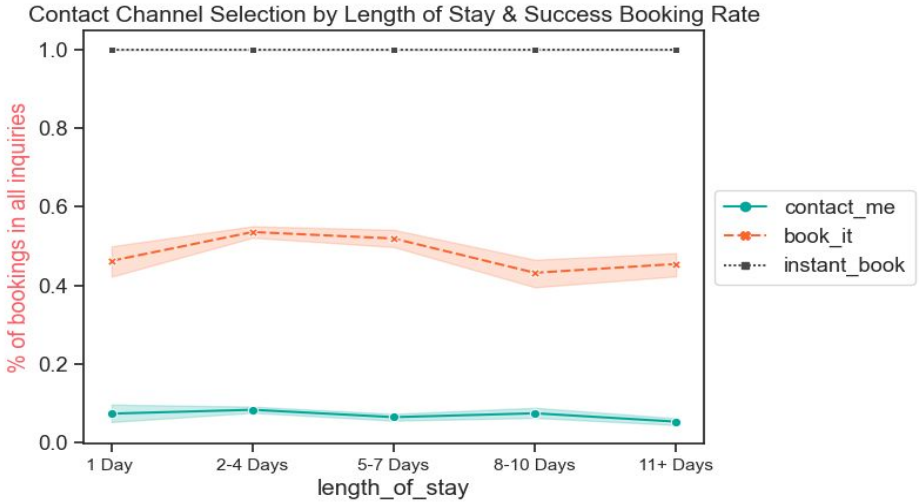




- Most of the inquiries are made for a group size between 1-2 people; as the group size increases, more users are likely to use the 'contact_me' channel and less likely to use the 'book_it' and the'instant_book' channels
- As the group size increases, the booking rate tends to deteriorate
- When the group size is greater than 6, the booking rate in the 'book_it' channel increases by 2%; however, the booking rate in the 'contact_me' channel decreases by 0.2%

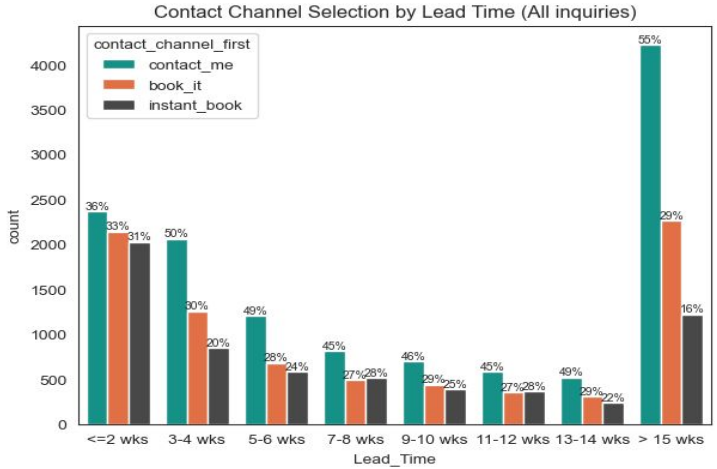
EDA - channel selection by length of stay & success rate

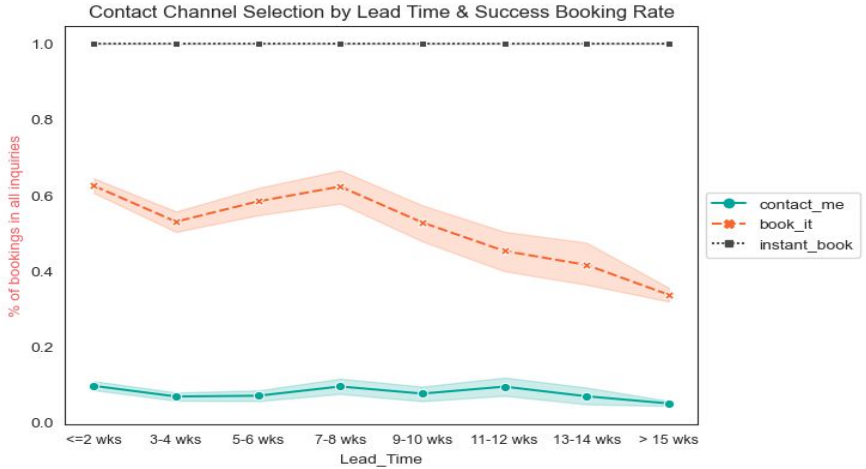




- Out of all inquiries, most of them are made for a 2-4 days trip
- The booking rate is the highest when the length of stay is between 2-4 days, and as the length of stay increases, the booking rate decreases gradually
- Amongst the inquires for a 1 day trip, there is no preference among the three channels; When the length of stay is greater than 8 days, more than half of the users would use the ‘contact_me’ channel
- When the length of stay is between 2-4 days, the booking rate is 54% for the ‘book it’ channel, which is six times higher than that of the ‘contact me’ channel

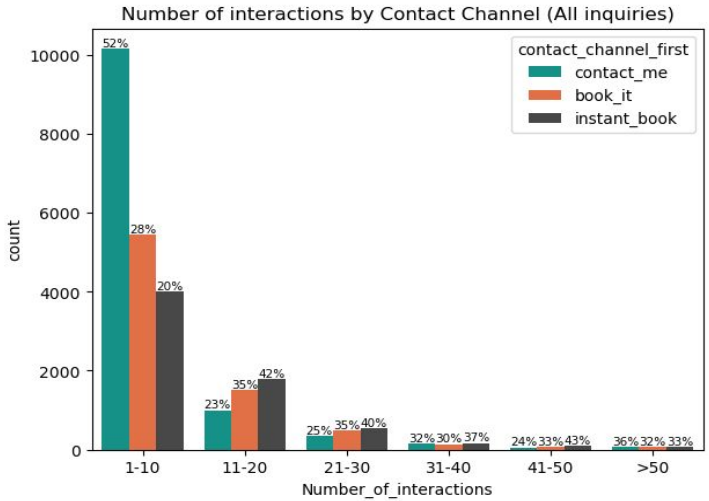
EDA - channel selection by lead time & success rate

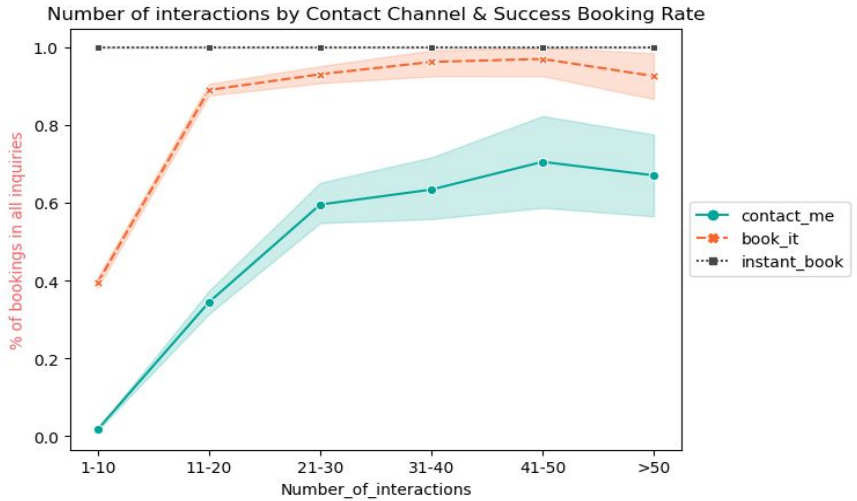




- There is no apparent preference for contact channel when the lead time is less than two weeks. As the lead time increase, more bookers are likely to use the ‘contact me’ channel and less unlikely to use the ‘instant book’ channel
- For inquiries with lead time between 0-2 weeks and 7-8 weeks, the booking rate is the highest in ‘book it’ and ‘contact me’ channel
- when the lead time <= 8 weeks, the booking rate in the ‘book it’ channel is about 6 time higher than that of the ‘contact me’ channel. The gap between the two channels gradually shrinks as the lead time > 8 weeks

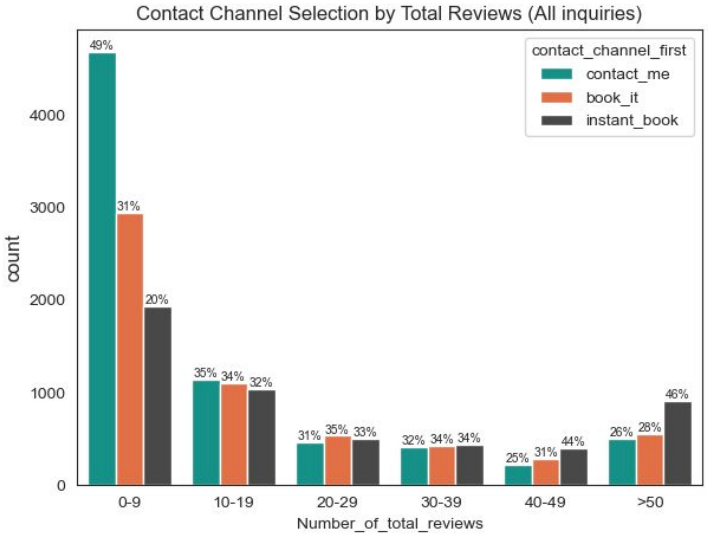
EDA - Number of interactions by contact channel & success rate

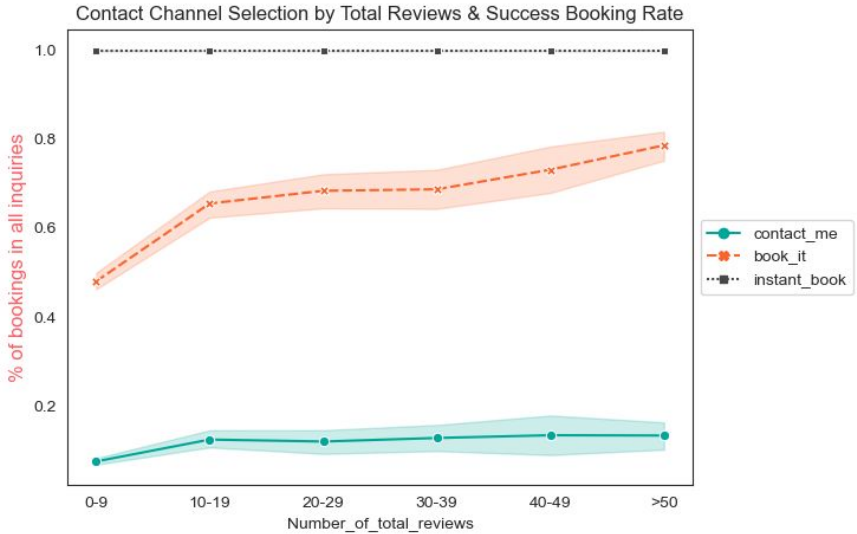




- Most of the inquiries have less than ten interactions
- The booking rate in both the ‘book it’ and the ‘contact me’ channels increased as the number of interactions rises
- As the number of interactions increased from 1-10 to 11-20, the booking rate in the ‘book it’ increased from 39% to 89%. And as the number of interactions increased from 11-20 to 41-50, the booking rate only grew 8%
- The booking rate increased 58% in the ‘contact me’ channel when the number of interactions increased from 1-10 to 21-30

EDA - channel selection by number of reviews per listings & success rate

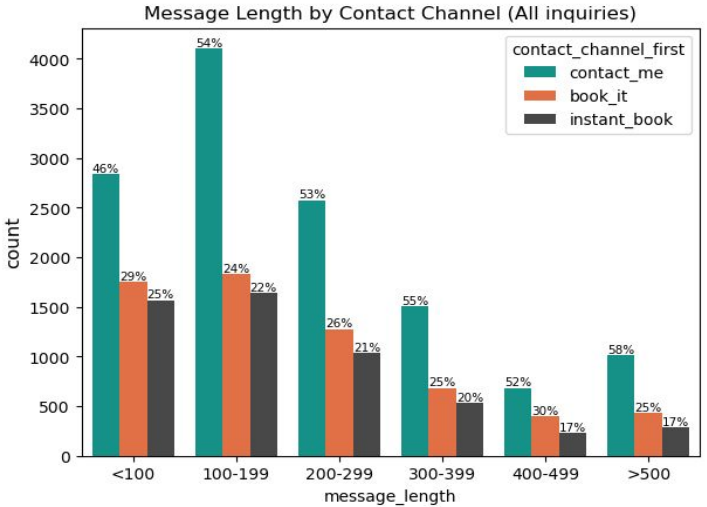


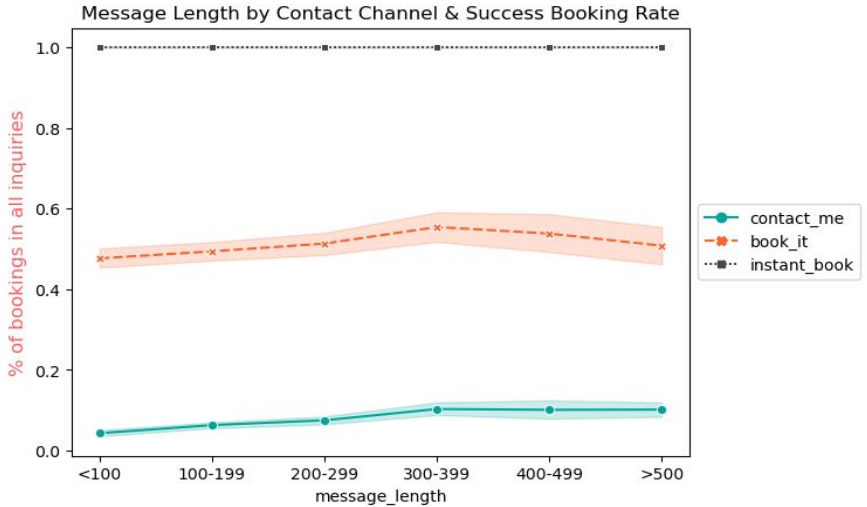


'book it'

- Of the listings with more than 50 reviews, 46% used the 'instant book' channel. However, for listings with less than 10 reviews, only 20% of users choose the 'instant book' channel
- When the number of reviews is greater than 50, the booking rate in the 'Book it' channel is 30% higher than those with 0-9 reviews

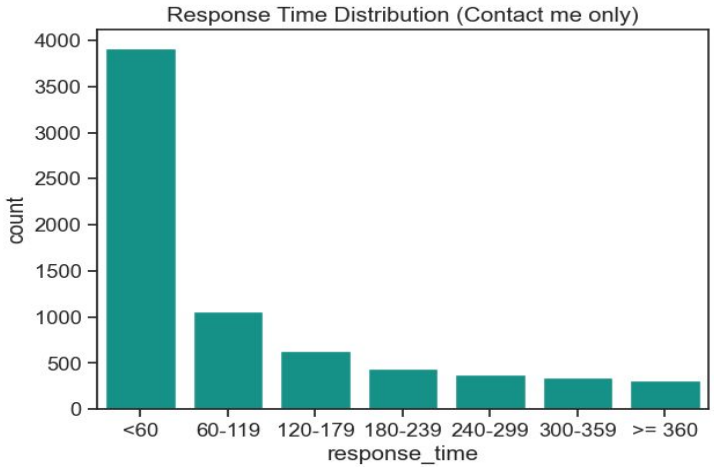
EDA - Message Length by contact channel & success rate

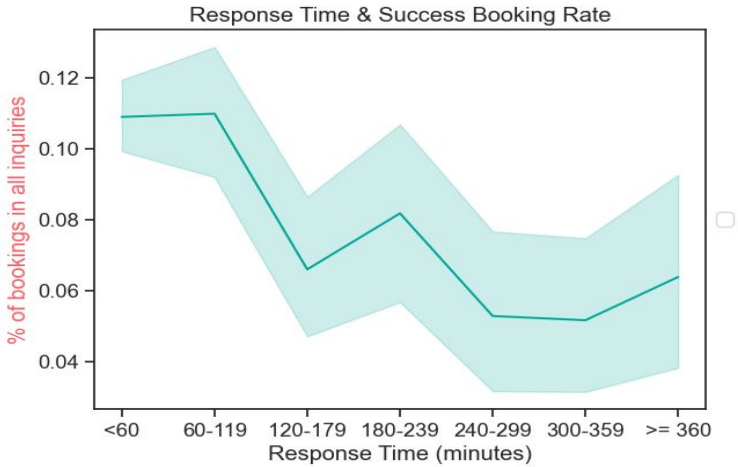




- Most inquiries message is between 100-199 characters.
- Inquiries with message length between 300-399 have the highest booking rate in both the 'book it' and 'contact me' channel; however, the booking rate in the 'book it' channel decreases as the message length > 400
- The booking rate increased 5% in the 'contact me' channel when the message length increased from less than 100 to greater than 500

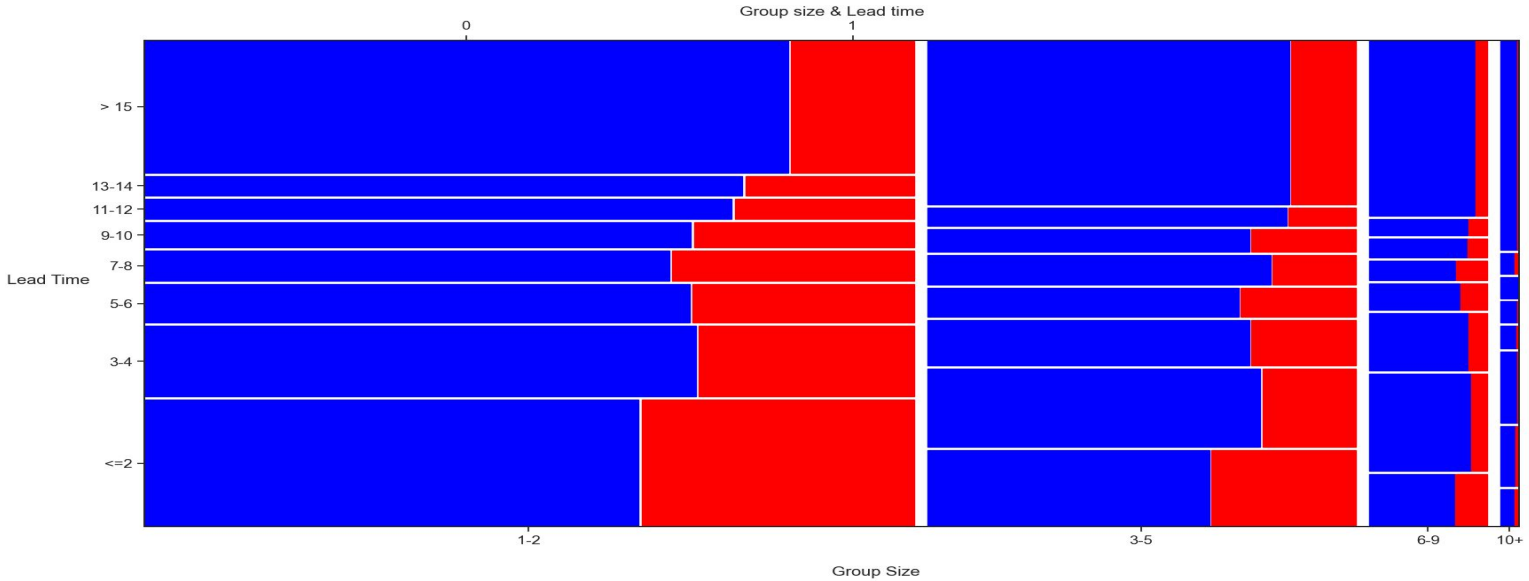
EDA - Host Response time by contact channel & success rate





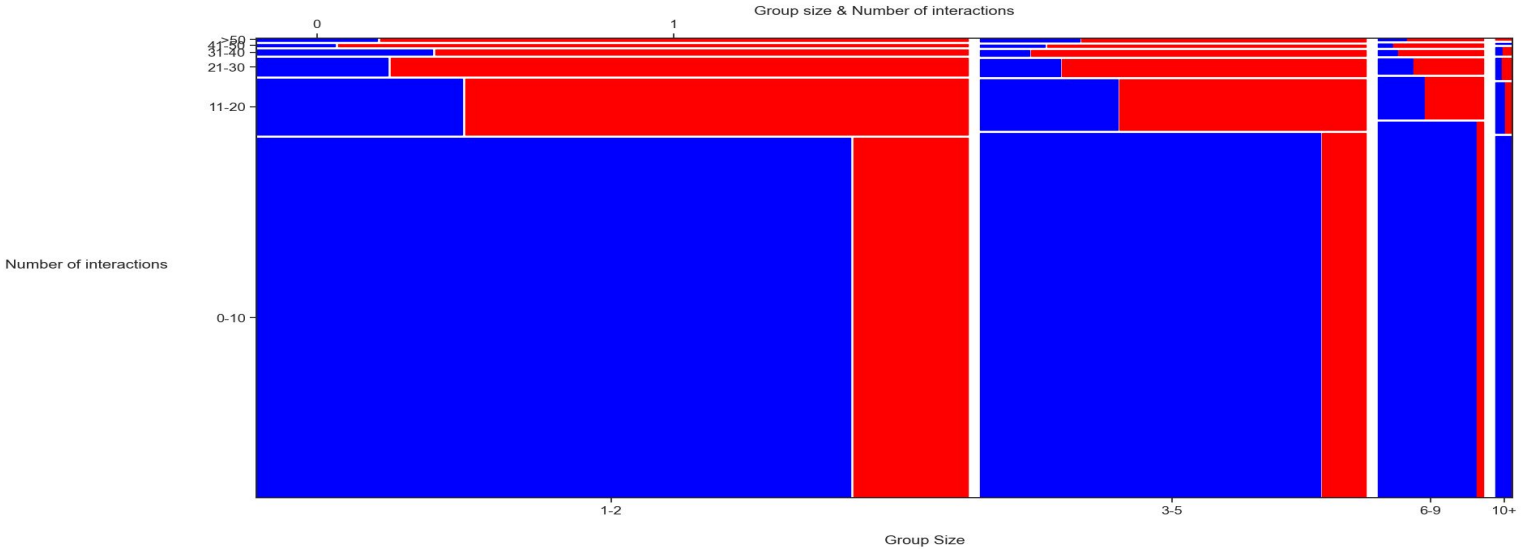
- Most of the inquiries in the ‘contact me’ channel got replied by the host within 60 minutes
- Inquiries got replied by the hosts between 0-59 minutes and 60-119 minutes have the highest booking rate, and the booking rate decreases as the response time increases

EDA - Mosaic Plot



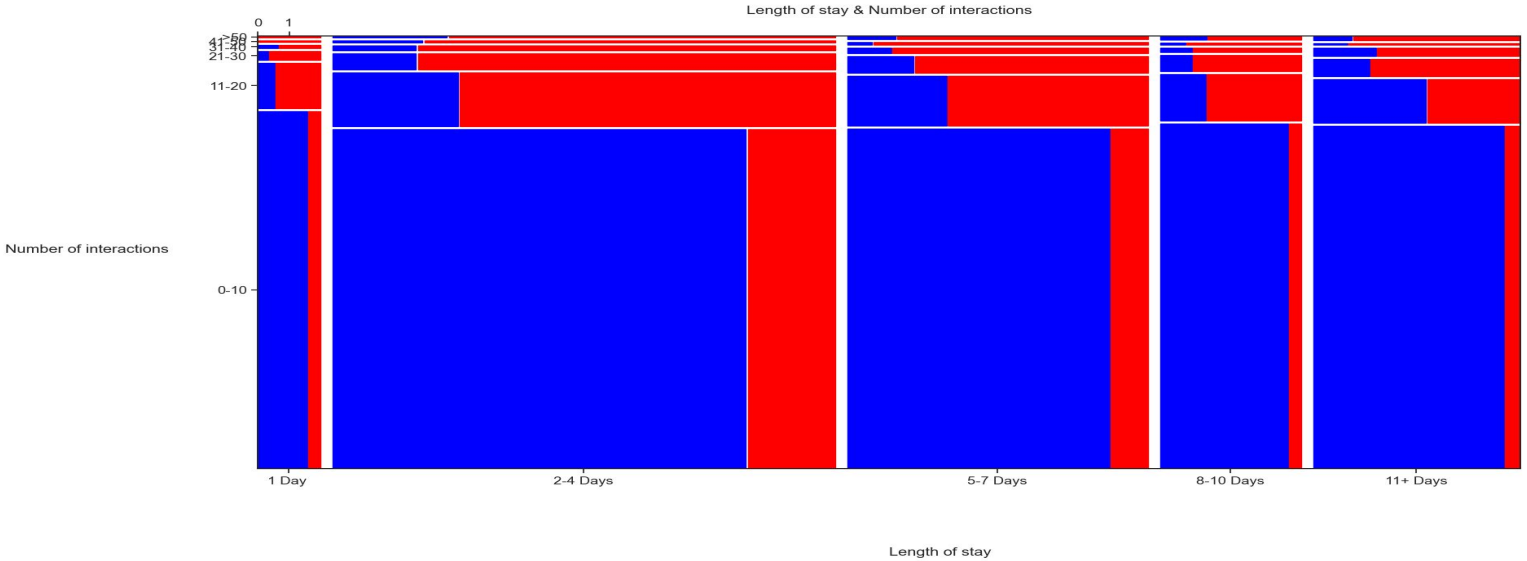
- In each group size bucket, the success booking rate is highest with a lead time <= 2 weeks.
- When the lead time >15, it can negatively impact the success booking rate in each group size bucket.
- When the group size is between 1-2, a lead time <=2 weeks give us the highest success rate, followed by a lead time between 7-8 weeks.
- When the group size increases from 1-2 to 3-5, increasing the lead time from 7-8 weeks to 9-10 weeks can help with successful booking.

EDA - Mosaic Plot



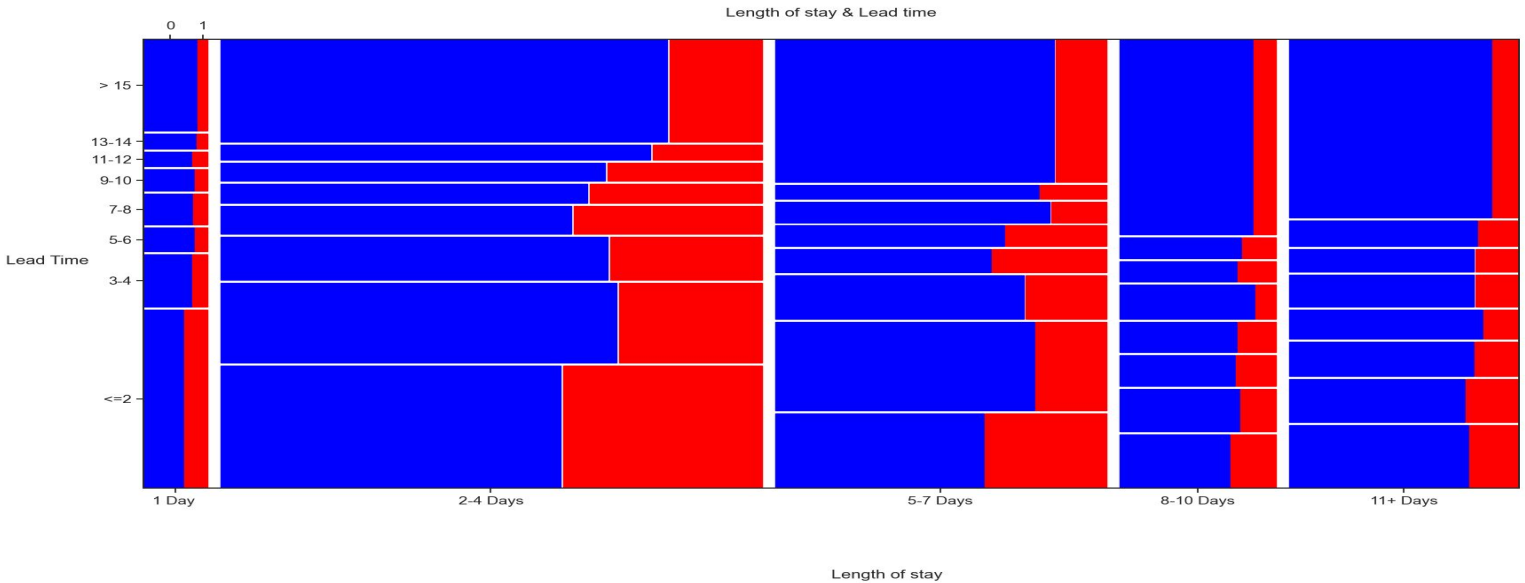
- The success booking rate increases as the interaction time increase from 0-10 to 21-30 regardless of the group sizet; however, when the interaction time increases to 31-40, the success booking rate change varies in each group size bucket.
- For each group size, increasing the interactions from 0-10 to 11-21 can yields the highest percentage increase in the success booking rate.
- When the number of interactions >50, it can negatively impact the success booking rate in each group size bucket.
- When the group size is between 1-2, interaction time between 41-50 yields the highest success booking rate and followed by interaction time between 21-30.

EDA - Mosaic Plot



- For each length of stay bucket, increasing the interactions from 0-10 to 11-21 can yield the highest percentage increase in the success booking rate.
- The success booking rate increases as the interaction time increase from 0-10 to 21-30 in all length of stay bucket; however, when the interaction time increases to 31-40, the success booking rate change varies in each length of stay buckets.
- When the length of stay is one day, the success booking rate is almost 1 when the number of interactions >=41.
- For the length of stay buckets between 2-4 days to 11+ days, when the number of interactions>50, it negatively impacts the success booking rate.

EDA - Mosaic Plot



- Regardless of the length of stay, the success booking rate is the highest when the lead time <= 2 weeks and lowest when the lead time >= 15 weeks.
- When the length of stay is one day, the success booking rate decreases gradually as the lead time rises.
- When the length of stay is between 2-4 and 5-7 days, the success booking rate drops as the lead time increases from <=2 weeks to 3-4 weeks. And the success booking rate increases afterward gradually and reaches the second highest success booking rate when the lead time is between 7-8 weeks.
- When the length of stay is greater than 11 days, the change in the success booking rate is trivial with a lead time <= 14 weeks; but when the lead time >= 15 weeks, it negatively impacts the success booking rate.