Expedia Hotel Bookings

Joseph van Bemmelen

Overview

- Project Background
- Data Overview
- Subsetting the Data
- Exploratory Data Analysis
    - → Business Intelligence
- Clustering
    - → Business Implications
- Groundwork for Recommendation Engine
- Future Steps

Project Background

- Expedia is a full-service online travel brand that allows customers to book hotels, cars, flights, cruises, and other vacation products

- The company performs hundreds of billions of predictive calculations annually to inform marketing decisions driving direct selling & marketing spend of ~$2.7B[1]

- More important than creating a completely accurate model, is using models to complement business intelligence and make an impact on the business

[1] Investor presentation, March 2016 (page 7)

Data Overview

- The data used for this project is repurposed from data made public by Expedia for a Kaggle competition (ended in 2013), which was focused on the position of a hotel in search results.

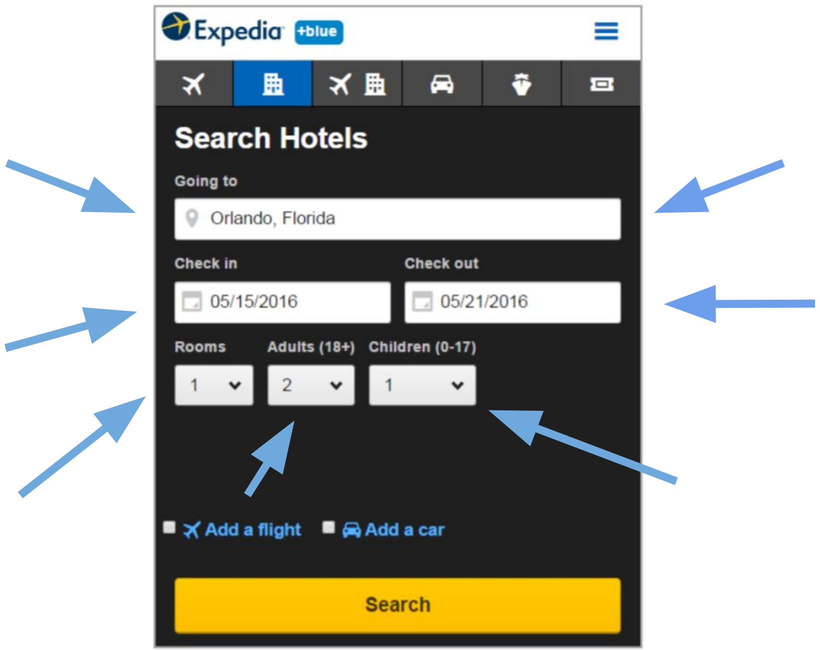
- Data is randomly sampled, however, converting impressions were oversampled from random impressions

- This project will analyze behavior of hotel trips, not necessarily users

- Dataset dimensions: 10 million rows, representing searches for 400,000 trips
  54 columns, representing many search characteristics

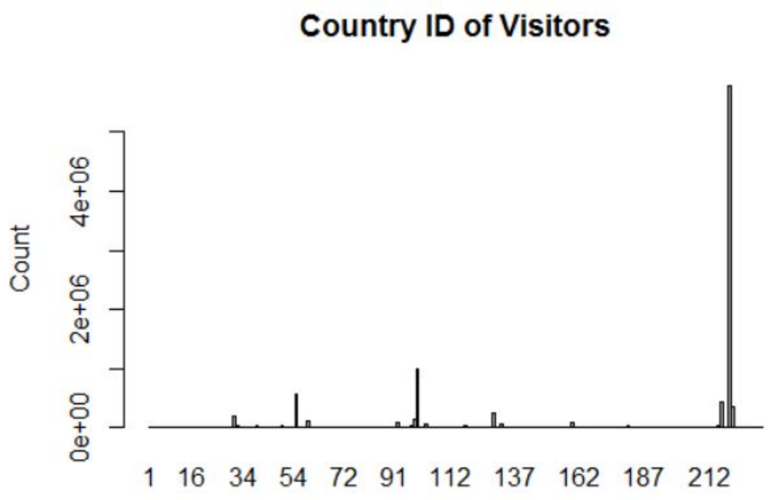
Data Overview

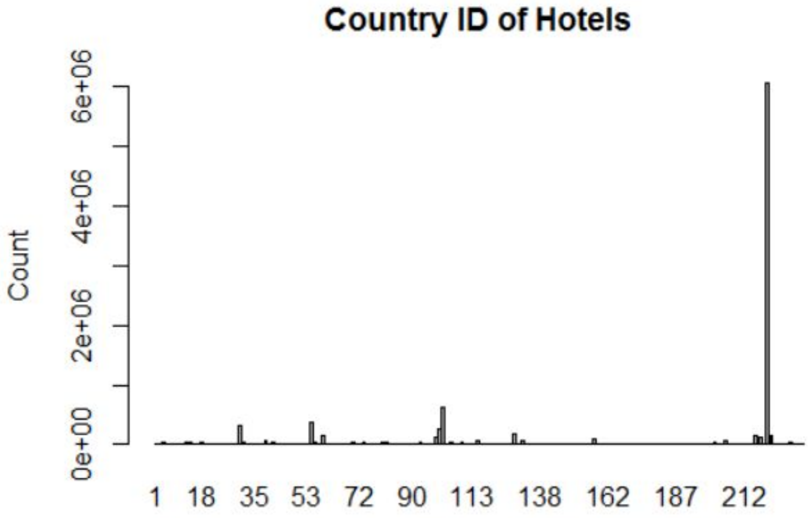
- Columns analyzed in this project:
    - Search ID
    - Country of user
    - Destination country
    - User’s historical ADR (average daily rate)
    - User’s historical hotel star rating
    - Position of property in search results
    - Whether position was random (boolean)
    - Number of rooms in search
    - Number of children
    - Number of adults
    - Length of stay
    - Booked or clicked result (both boolean)



Subsetting the Data

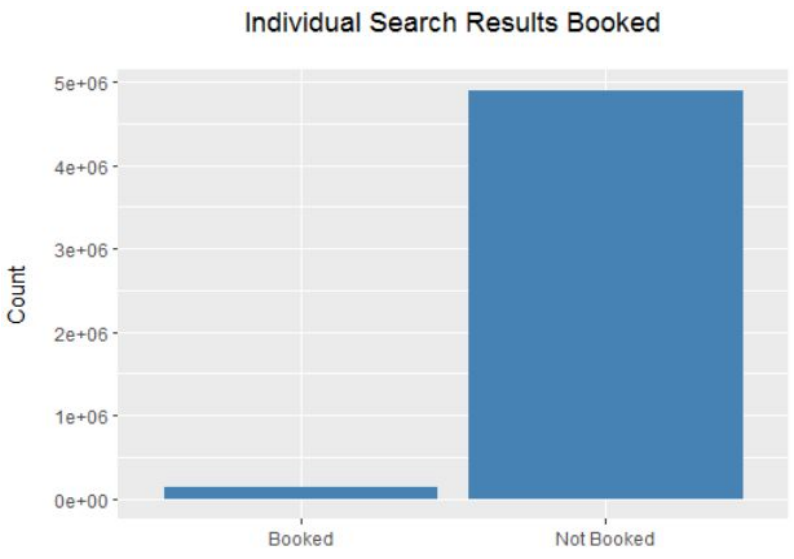
- Country 219 has the vast majority of visitors and hotels, we will focus on this subset which we will assume is the US (5M rows, 200,000 trips)

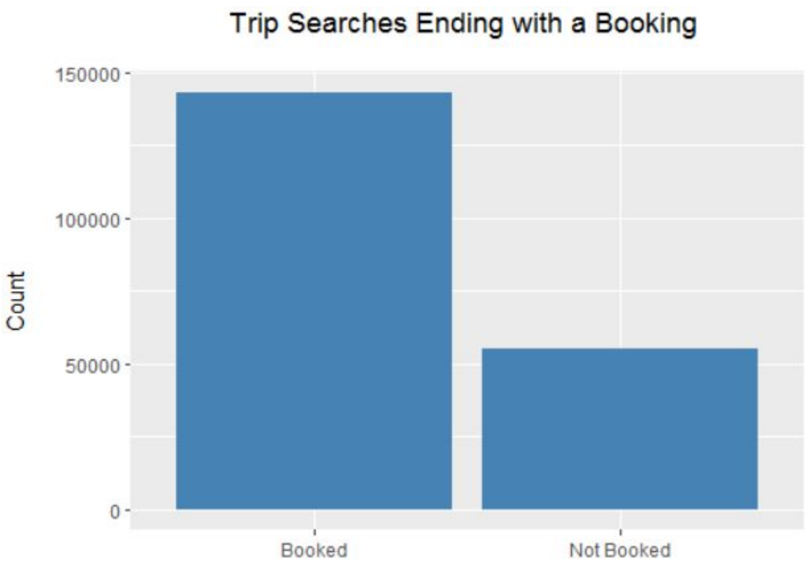




Data Overview

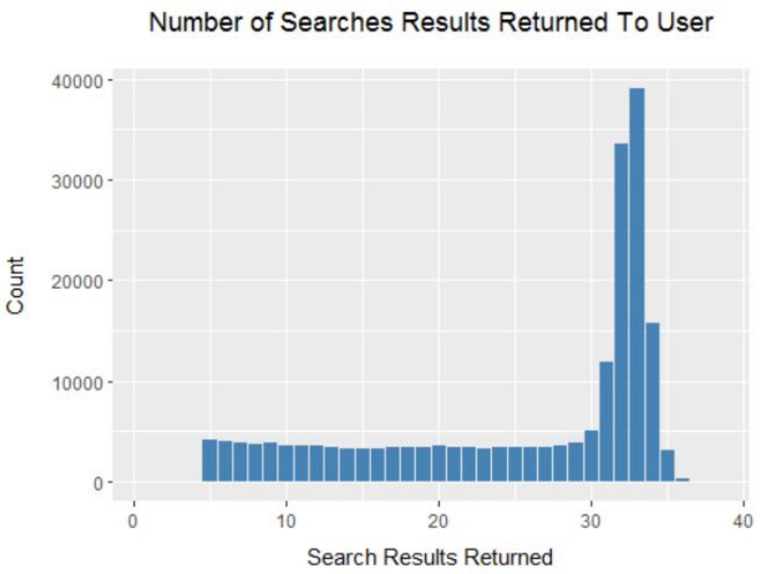
- Most search results are not booked, however, converting impressions were oversampled so most trip searches do end with a booking

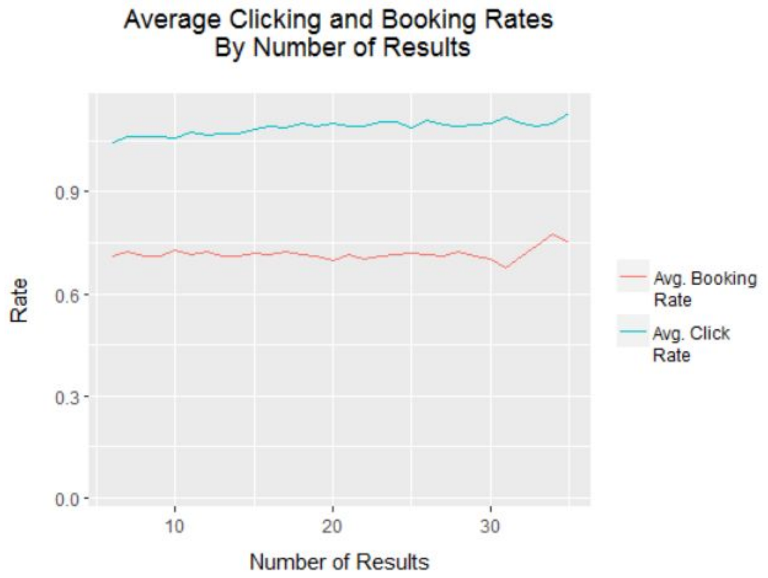




Data Overview

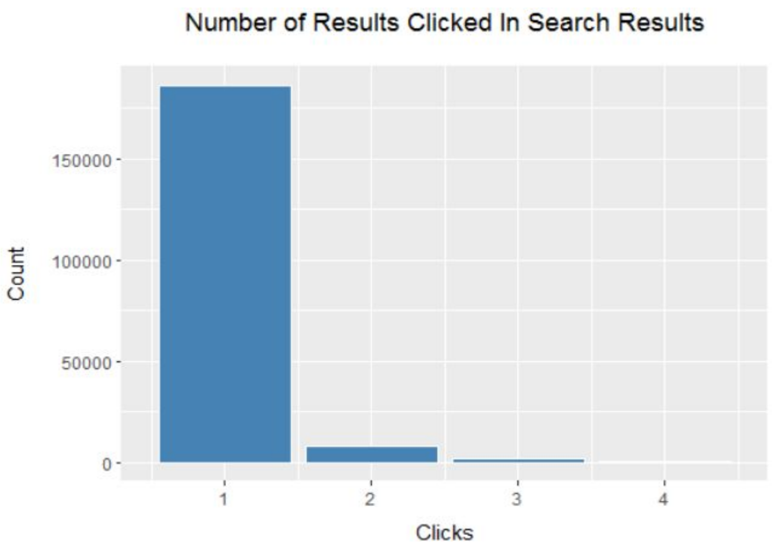
- Searches often have 30-35 results returned to the user
- Clicking and booking rates relatively stable across number of results returned





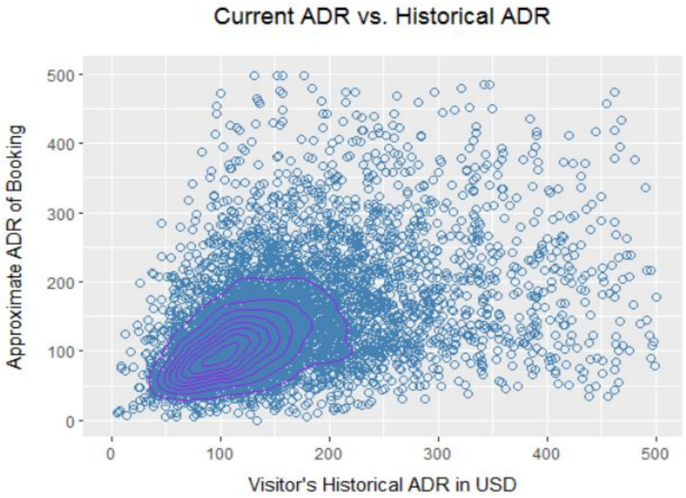
Booking Rate by Clicks

- The booking rate appears to decline as the number of clicks increases (indicative of comparison shopping)



Bookings Comparable to Historical Preferences

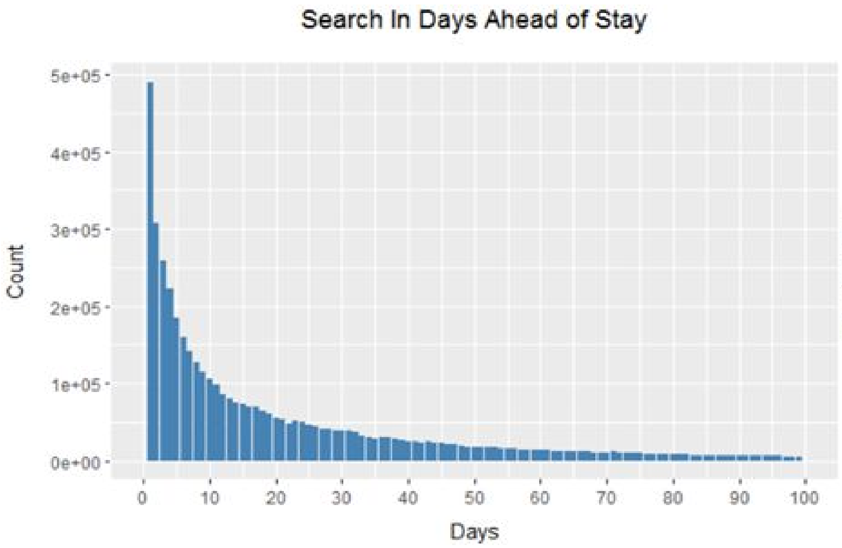
- Current booking is generally in line with a visitor's past booking preferences
- However, historical data only available for 4.0% of trip searches (MAR, bias)



[1] Current ADR may include taxes taxes, fees, conventions on multiple day bookings and is manually calculated as (Gross Booking in USD / Room Count) * (1 / Length of Stay)

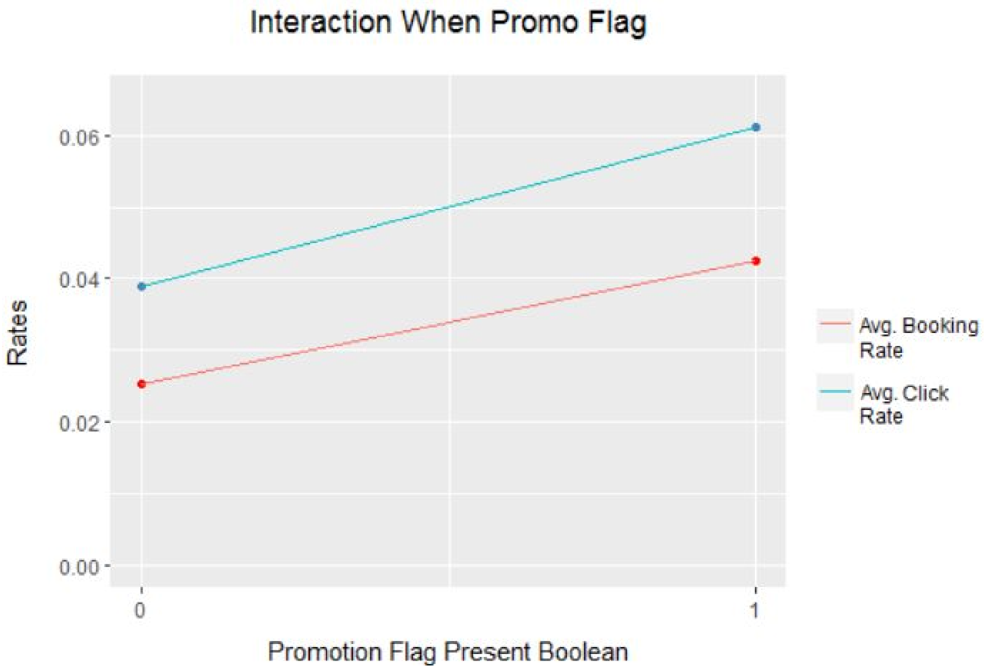
Booking Rate by Window

- The booking rate declines as the number of days ahead of stay increases (indicative of browsing/comparison shopping)



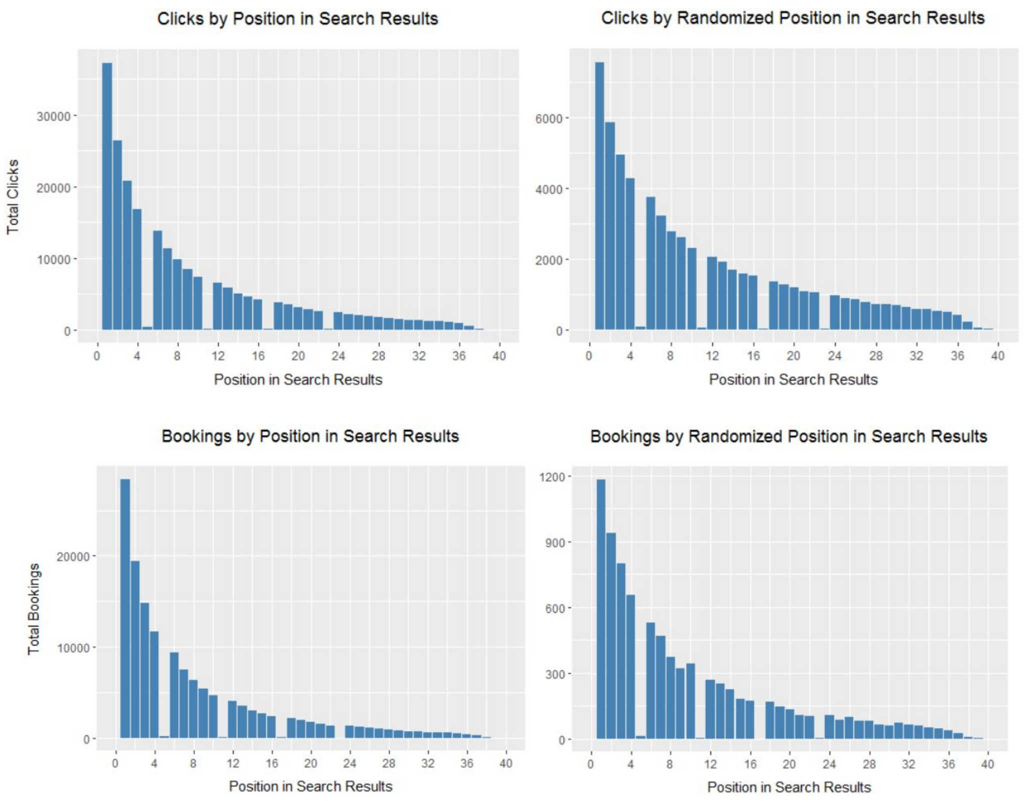
Impact of Promo Flag

- As expected, there are more clicks and bookings when promo flag present



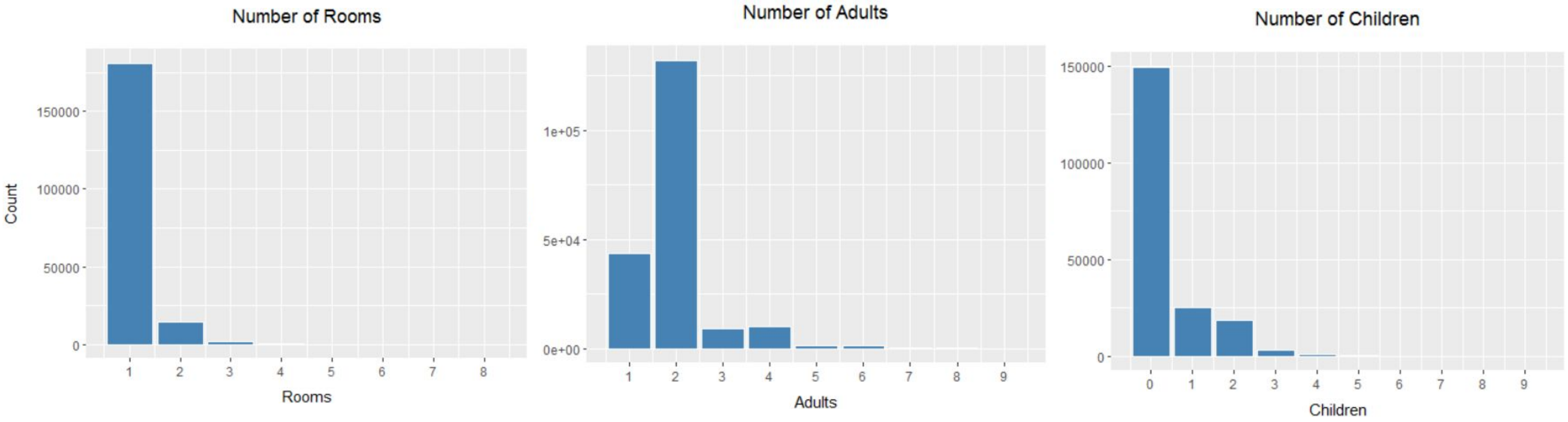
Impact of Order in Search Results

- There are more clicks and bookings when promo flag present, even when randomized order
- Charts in the right column only contain clicks for search results returned in a randomized position
- Introduced bidding for prime placement early 2016



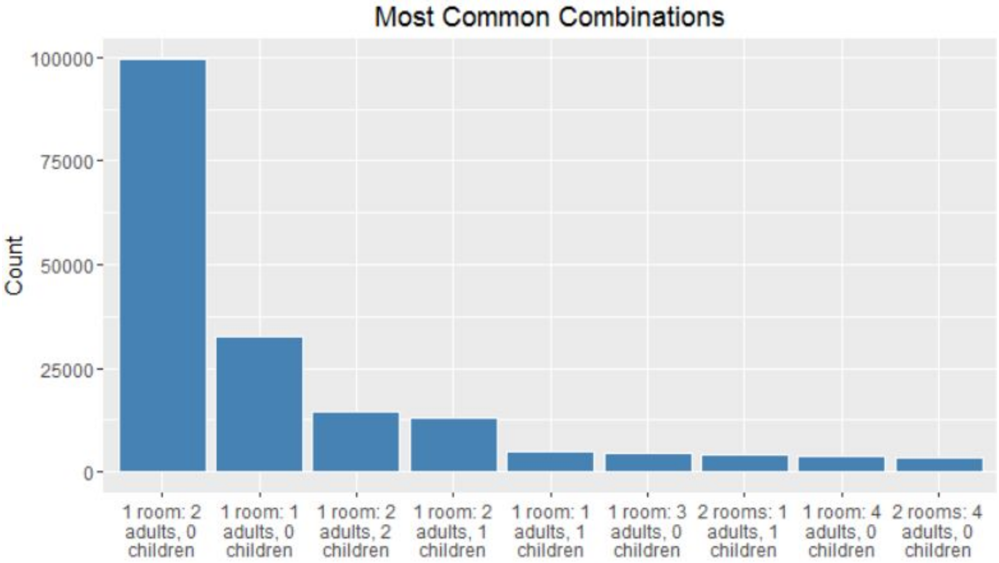
Most Common Searches

- The most common search is for one room, two adults, and no children



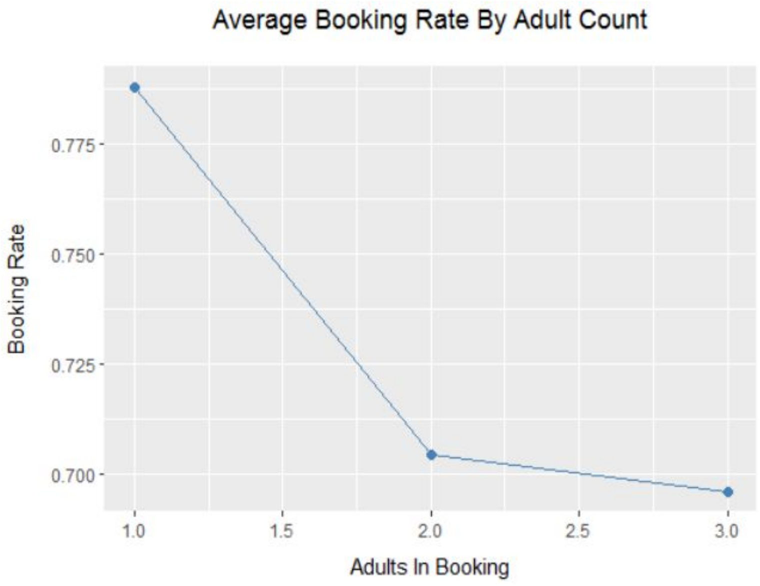
Most Common Search Combinations

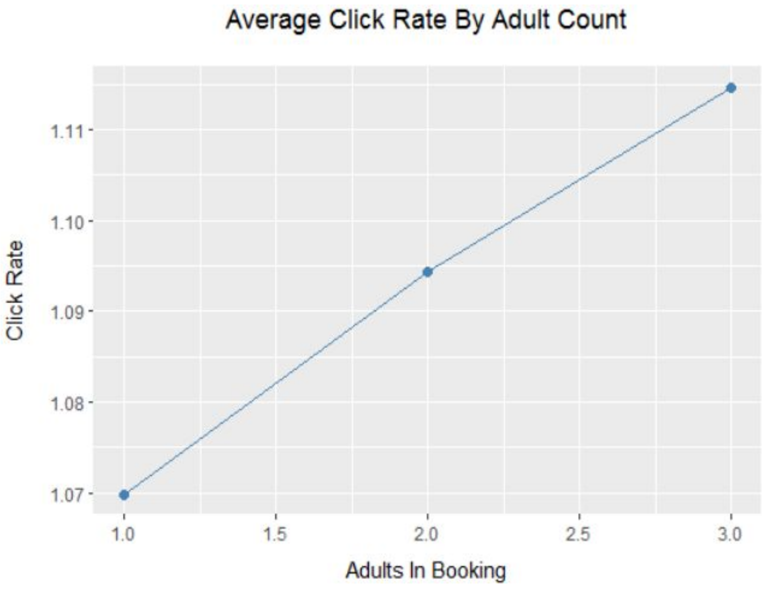
- Most common booking is for two adults with no children, then one adult with no children



Adults in Booking Trends

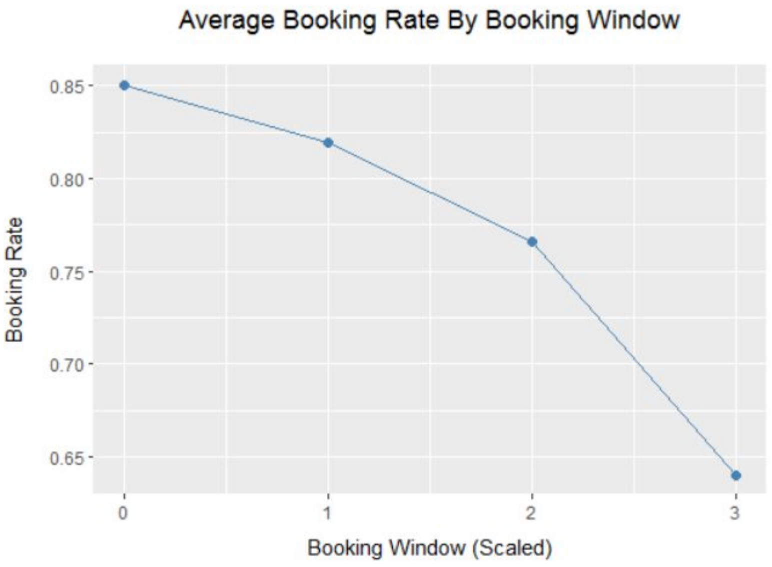
- More adults have a higher click rate, but lower book rate
  (adults scale: 1 = 1, 2 = 2, 3 = 3 or more adults)

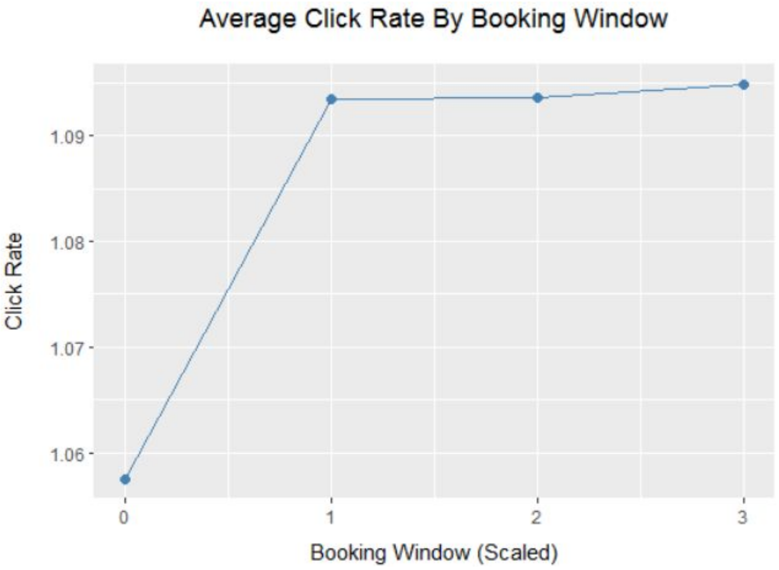




Booking Window Trends

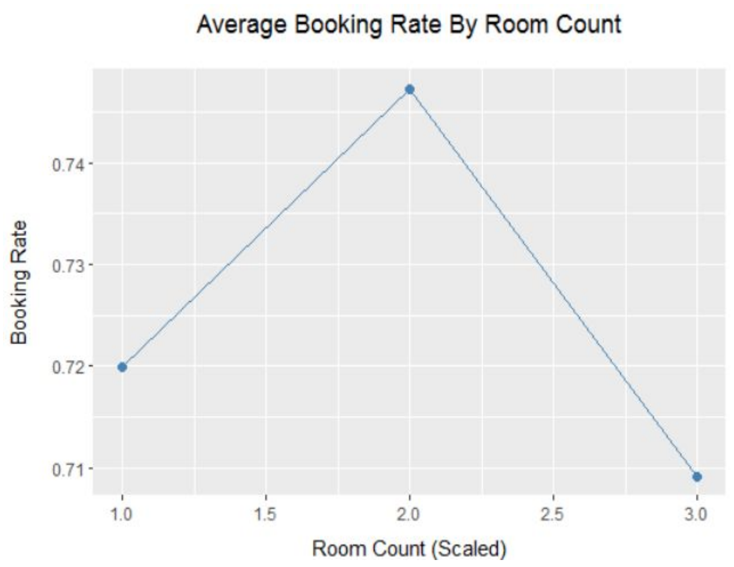
- Searches farther in advance have a higher click rate, but lower book rate
  (booking window scale: 0 = 0, 1 = 1, 2 = up to 2 weeks, 3 = over two weeks)

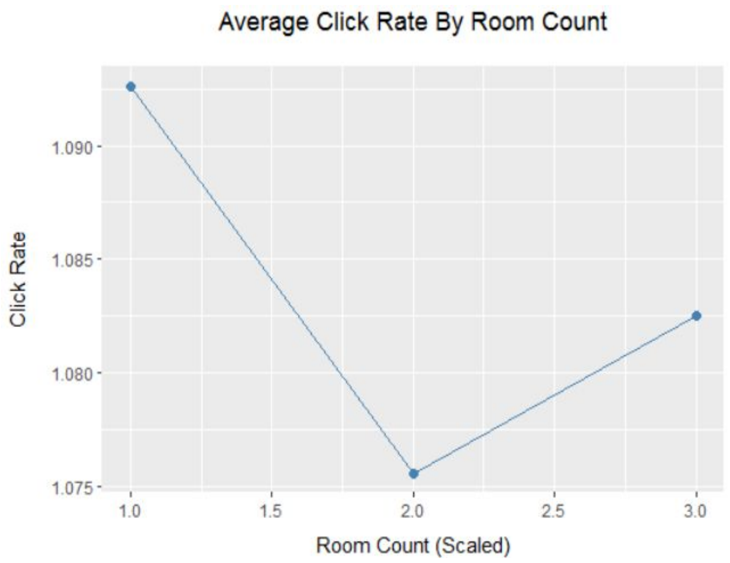




Room Count Trends

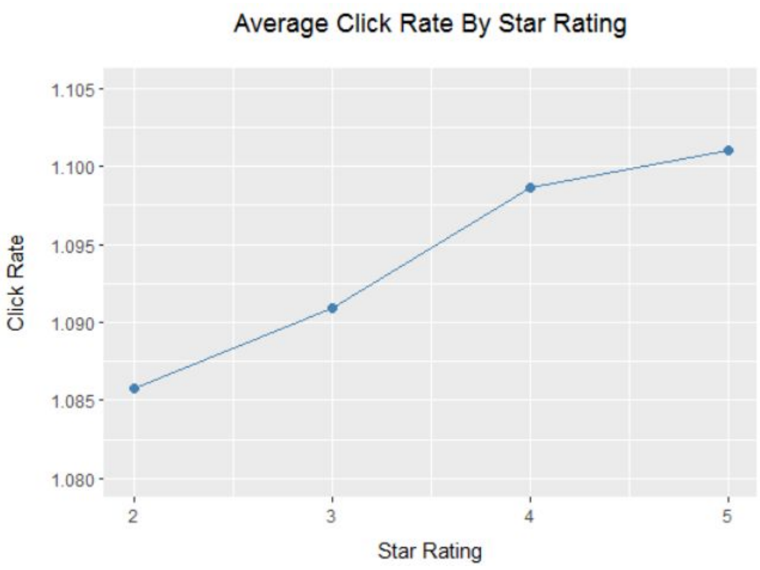
- Bookings and click rate by room count are inverse
  (room count scale: 1 = 1, 2 = 2, 3 = 3 or more rooms)





Hotel Stars Trends

- Users more likely to click when a hotel has more stars and more likely to book, except for 5-star hotels (look, but not book) [3.0% missing star-rating]



Clustering Trips

- Clustered on number of rooms, number of adults, number of children, booking window, whether the trip included a Saturday night, and length of stay
- Using K-means and Euclidean distance measurement
- Manually adjusted each measure and created different levels for continuous variables
- Used a matrix that compared the figures of each cluster to the mean in order to find clusters with the largest spreads
- Iterative process
- Three main groups of travelers: business travelers, couples, and families

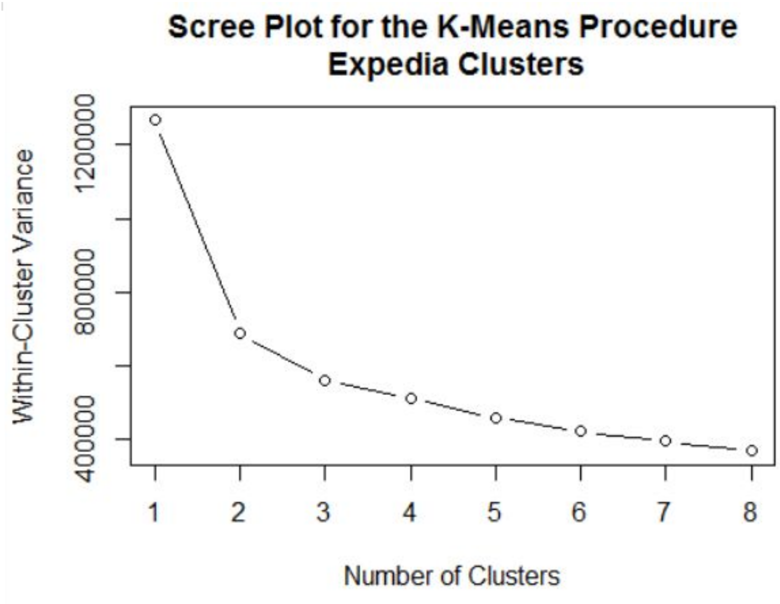
Clustering Trips

	Booking rate	Click rate	Hotel stars	Hotel price	Gross bookings	Historical hotel stars	Historical ADR
1	0.8040400	1.086077	2.919327	119.7945	206.9190	2.931868	145.0936
2	0.7283412	1.098922	3.646682	180.2624	418.8853	3.201598	182.9270
3	0.5945386	1.089377	2.345945	130.5406	305.3681	2.875510	148.5553

	Adults	Children	Rooms	Saturday stay	Length of stay	Booking window
1	1.266195	0.1336265	1.084484	0.5752883	1.579252	2.244071
2	1.688095	0.4118689	1.110355	0.5676018	2.506560	30.972495
3	2.189114	1.8836269	1.140161	0.5666025	2.236358	58.732979

Project Background

- There's a slight elbow at three on the scree plot as well, which minimizes the within-cluster variance, although this is not a deciding factor here



Business Takeaways

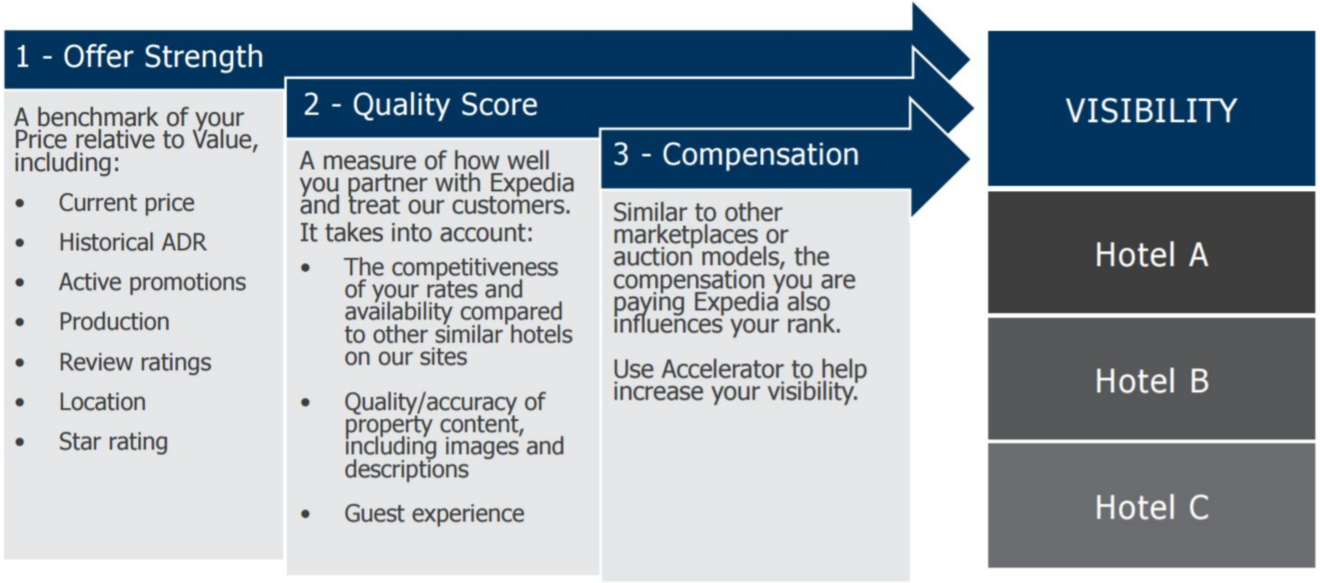
- When search inputs indicate that a trip is in cluster #2, there should be more resources aimed at acquiring that user because he has the highest gross bookings
- However, trips in cluster #1 have the highest booking rate and book with the shortest notice, might be best to convert them on the website using a promotion
- Trips in cluster #3 are usually at lower star hotels and have the lowest booking rate, may do more comparative shopping, emails closer to the trip might be a better approach

Groundwork for Recommendation System

- Additionally, each cluster would prefer slightly different hotels, so knowing the cluster can help in ordering the hotels and in providing relevant recommendations
- Recommendations for hotels have to be limited to the geographical location that the trip is in
- Using the trip clusters, we can find which hotels in a location were booked the most often by those in the same cluster and find the hotels that have the highest average reviews

Groundwork for Recommendation System

- This would be an improvement over the company's current ranking algorithm, which does not take into account a recommendation system



[1] Expedia Marketplace White Paper (April 2016)

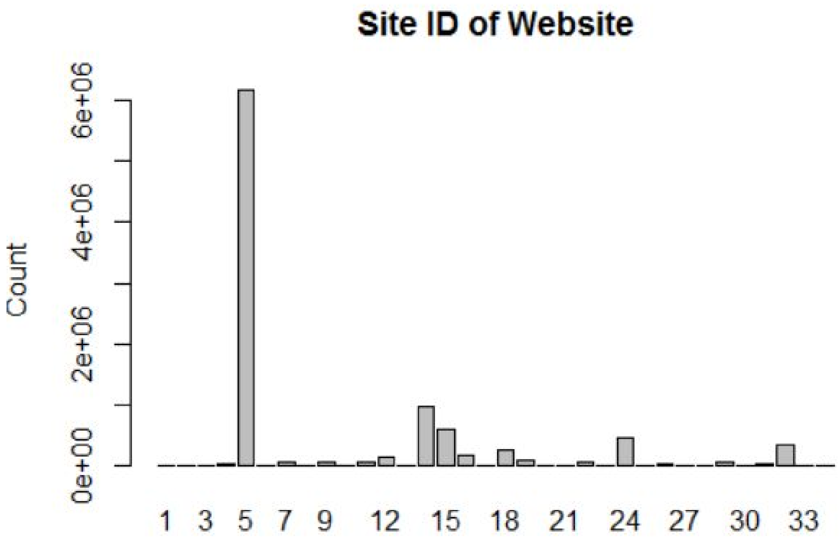
Future Steps

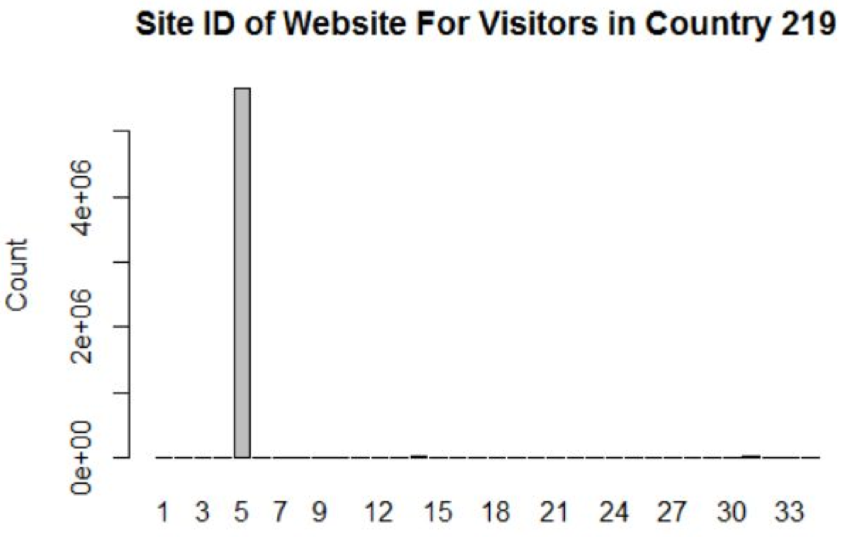
- Work on implementing recommendation system
- Continue to refine clusters and potentially identify additional traveller groups
- Expand analysis to other countries in the Expedia data
- Compare to purchase behavior of other websites

Thank you!

Subsetting the Data

- Similarly, site ID labelled “5” has the majority of overall visitors and the vast majority of visitors from country 219



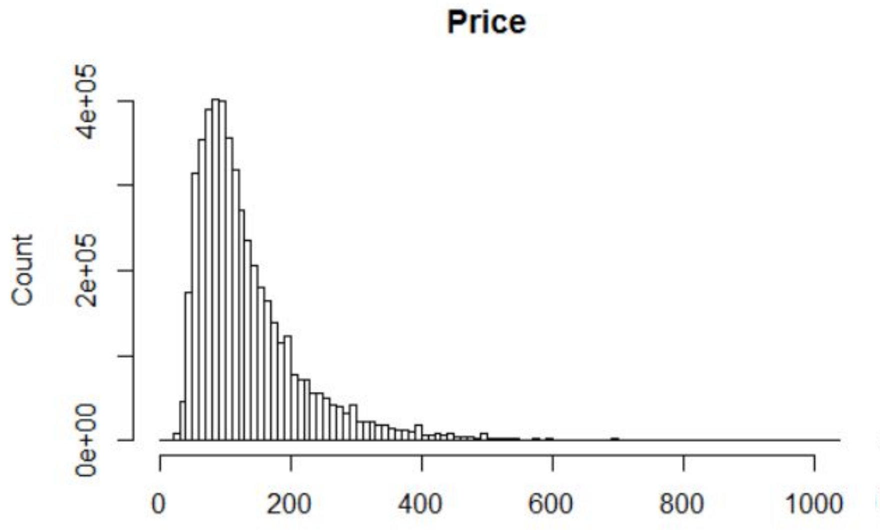


Goal: Increasing Gross Bookings

- Expedia doesn't profit from flights and instead profits from hotels, rental cars
- Gross bookings is the figure we aim to maximize, which may include taxes/fees

Project Background

Price has high outliers - remove since fake? Then do log transformation?



Article Frequency