Airbnb New User Bookings

Which user will make his or her first booking?

by
Sarah Huang

Overview
The objective is to predict if a user would make a booking.

Data Files
A list of users along with their demographics
Training Set: 213,451 observations of 16 variables
Testing Set: 62,096 observations of 15 variables
Web session records 10,567,737 observations of 6 variables

Training/Testing Set
user id
the date of account creation
timestamp of the first activity
date of first booking
gender
age
signup method
signup flow: the page a user came to signup up from
language: international language preference
affiliate channe
affiliate provider
first affiliate tracked..

Web Session Record
user id
action (360 classes)
action type (11 classes)
action detail (156 classes)
device type (14 classes)
secs elapsed

Preprocessing and Feature Engineering

1. Data-Cleaning

2. Add features from date account created, and timestamp first active by extracting weekday, day, week number.

3. Add features from Sessions using combination of action, action type, action detail. The result is 457 combinations. The result is then trimmed by keeping combinations with more than 20 observations.
- Aggregation is done on the ID level + action combination, using number of occurances, flags, sum of seconds elapsed, average of second elapsed.
- Aggregation is also done on the ID level only, summing total number of sessions per user, total number of seconds elapsed per user, and average seconds elapsed per user.

4. One-hot encoding is used to code all factor variables.

5. Data is split into training and evaluation set with 1,654 variables.

6. Create two sets of predictors, the full set contains all predictors, the reduced set exclude predictors that are sparse, near zero variance, or highly correlated.

Modeling

R, Tableau, and the H2O Flow were utilized in the analysis.

- Two class classification model - distinguish between users that booked, and users that did not book.

- Models were run with both the fullset of predictors, reduced set of predictors (no nzv, no sparse, no highly correlated), as well as predictors corresponding to a set of keywords.

- Models were run with 10-fold cross validation (CV) and evaluated by accuracy, kappa, and exact matching error (merror)

- Models were tuned using caret package in R.

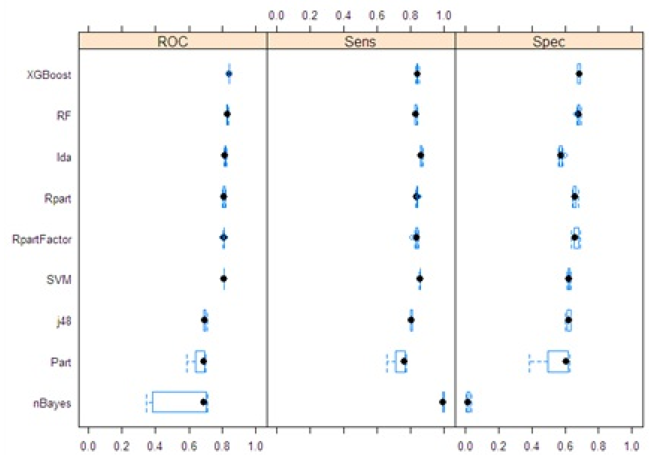
- A list of models were run and an ensemble was created with a combination of the 4 best performing models (random forest, neural network, rpart, xgboost)

- List of Models: Linear Discriminant Analysis (LDA), Partial List Square (PLS), Lasso and Elastic-Net Regularized Generalized Linear Models (GLMNET), Neural Network (NNET), Flexible Discriminant Analysis (FDA), K-Nearest Neighbors (KNN), Naive Bayes, Basic Classification Tree (Rpart), J48, Random Forest (RF), XGBoost (XGB)

- A separate ensemble was created using H2O (Base Learner: Generalized Linear Modeling (GLM) , Gradient Boosting Method (GBM), Distributed Random Forest(DRF) and stacked using Deep Learning & GLM.

Model Summary

Individual Model Performance:



Ensemble using H2O:

```
Base learner performance, sorted by specified metric:
                         learner       AUC
2 GLM_model_R_1476250914526_1419 0.8046919
3 DRF_model_R_1476250914526_2123 0.8228619
1 GBM_model_R_1476250914526_1444 0.8384450


H2O Ensemble Performance on <newdata>:
----------------
Family: binomial

Ensemble performance (AUC): 0.840592053553151
```

How is the result?

Airbnb New User Bookings

8 months ago · Top 18%

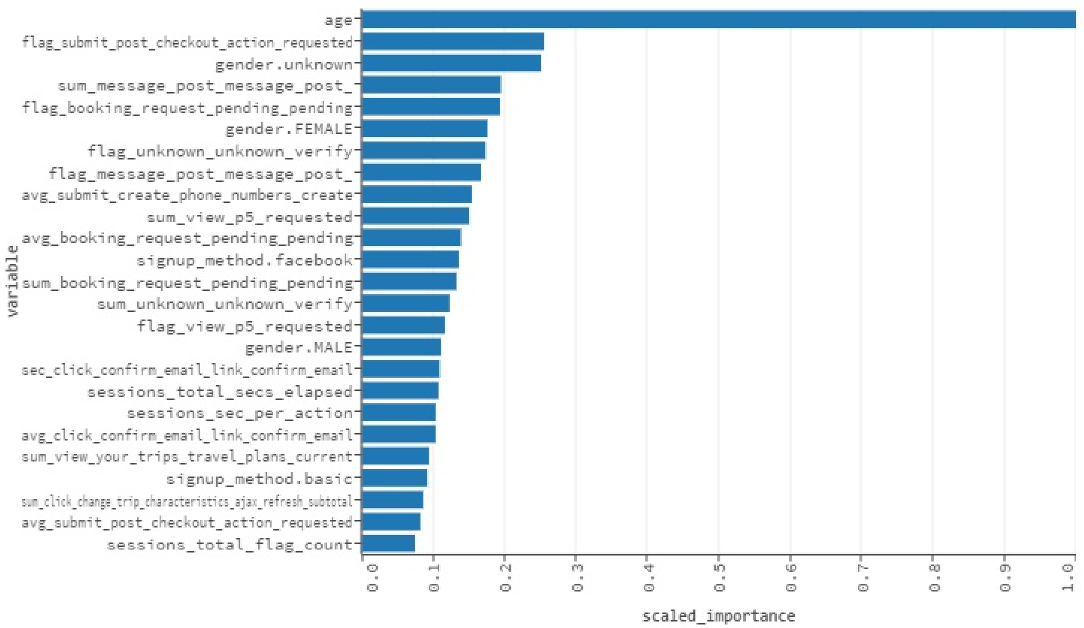
10 entries as a solo competitor

252nd
of 1462

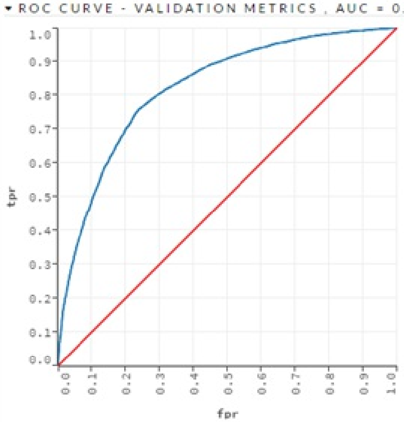
Pre-Processing | Modeling | Model Summary | Sample model ouptut | Data Exploration: Country | Data Exploration: Gender | Data Exploration: Device

Sample Model Run

▾ VARIABLE IMPORTANCES



▾ ROC CURVE - VALIDATION METRICS , AUC = 0.822862

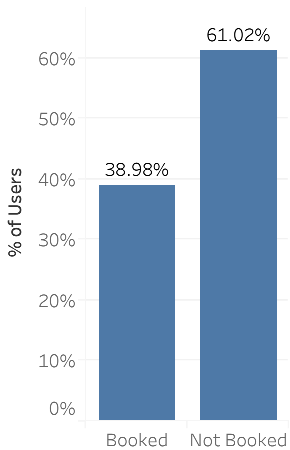


Modeling | Model Summary | Sample model ouptut | Data Exploration: Country | Data Exploration: Gender | Data Exploration: Device | Data Exploration: Affiliate (marketing channel)

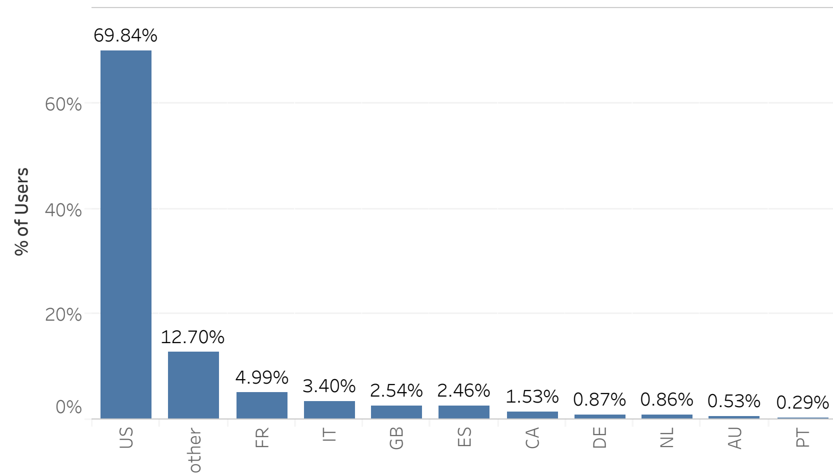
Which countries are popular amongst the travelers?

Of the 39% of users that made a booking, the most popular destination is US.

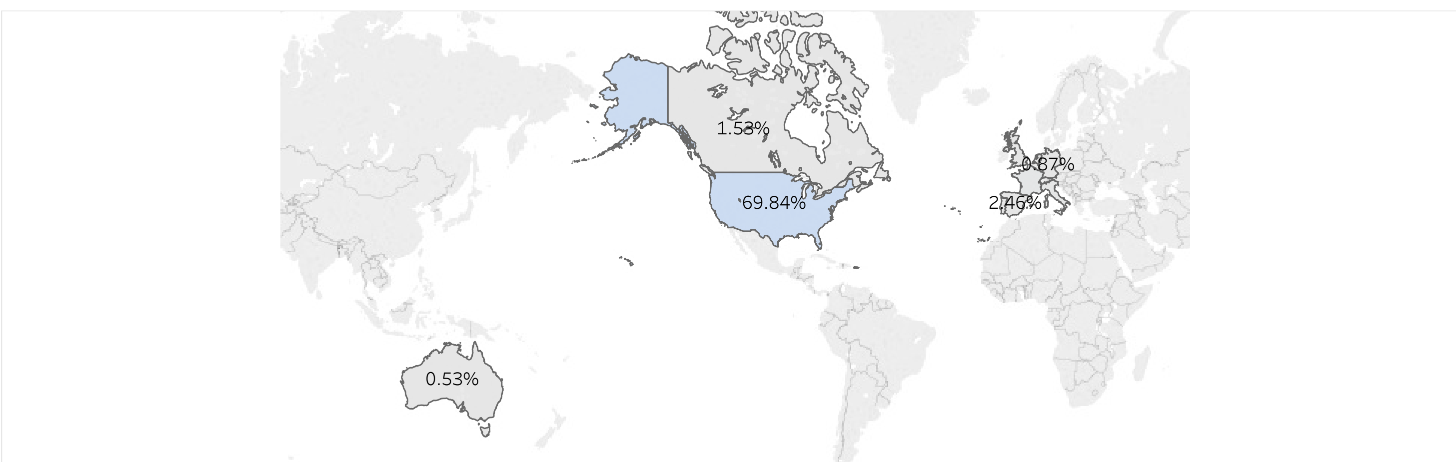
Booking %



Breakdown of Booking By Country



Destination Map

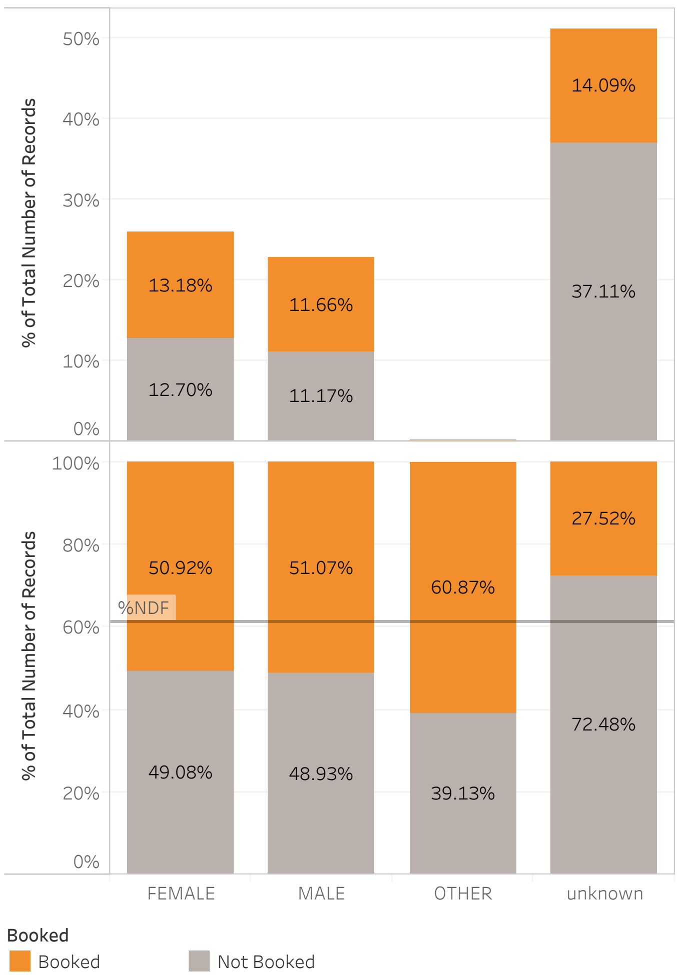


Model Summary | Sample model ouptut | Data Exploration: Country | Data Exploration: Gender | Data Exploration: Device | Data Exploration: Affiliate (marketing channel) | Data Exploration: Sign-up Method

Popularity of Destination by Gender

Females and males have similar booking ratio and similar country preferences. The booking ratio is lowest for the users with unknown gender.

Booking Ratio by Gender

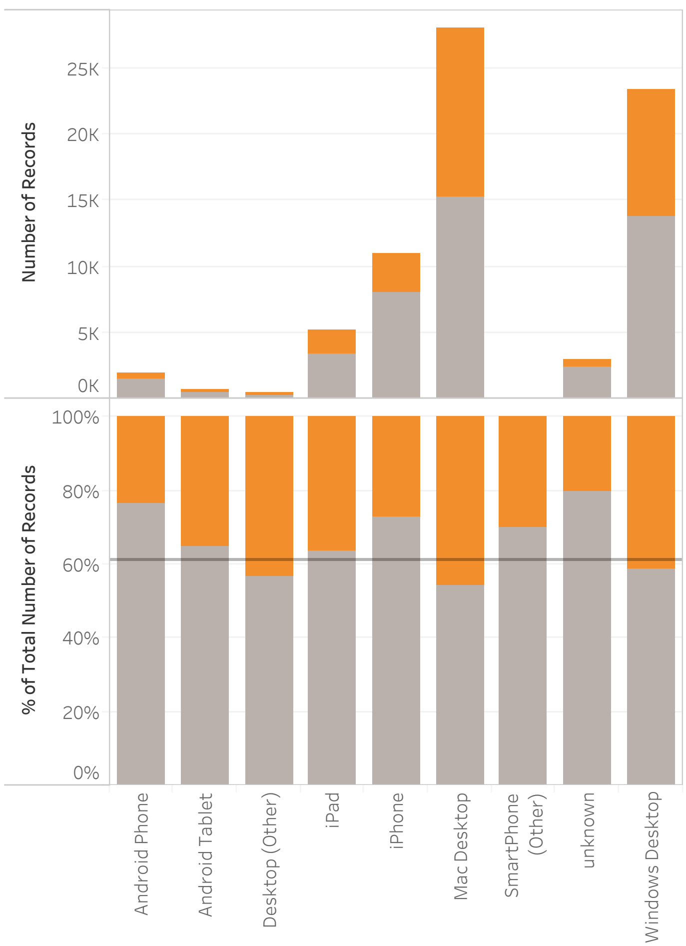


Destination by Gender

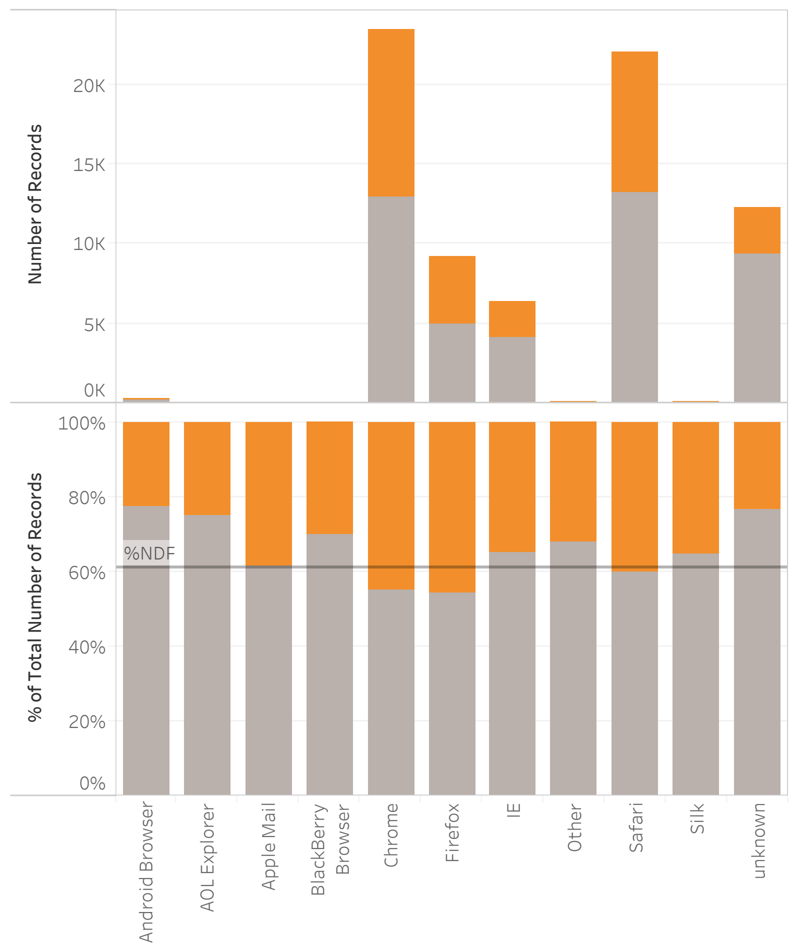
Which device and browser has higher percentage of users who booked?

Mac Desktop not only is the most popular first device, users who used it has the highest booking ratio.
Chrome is the most popular first browser, and was the 2nd highest in booking ratio.

Device Type



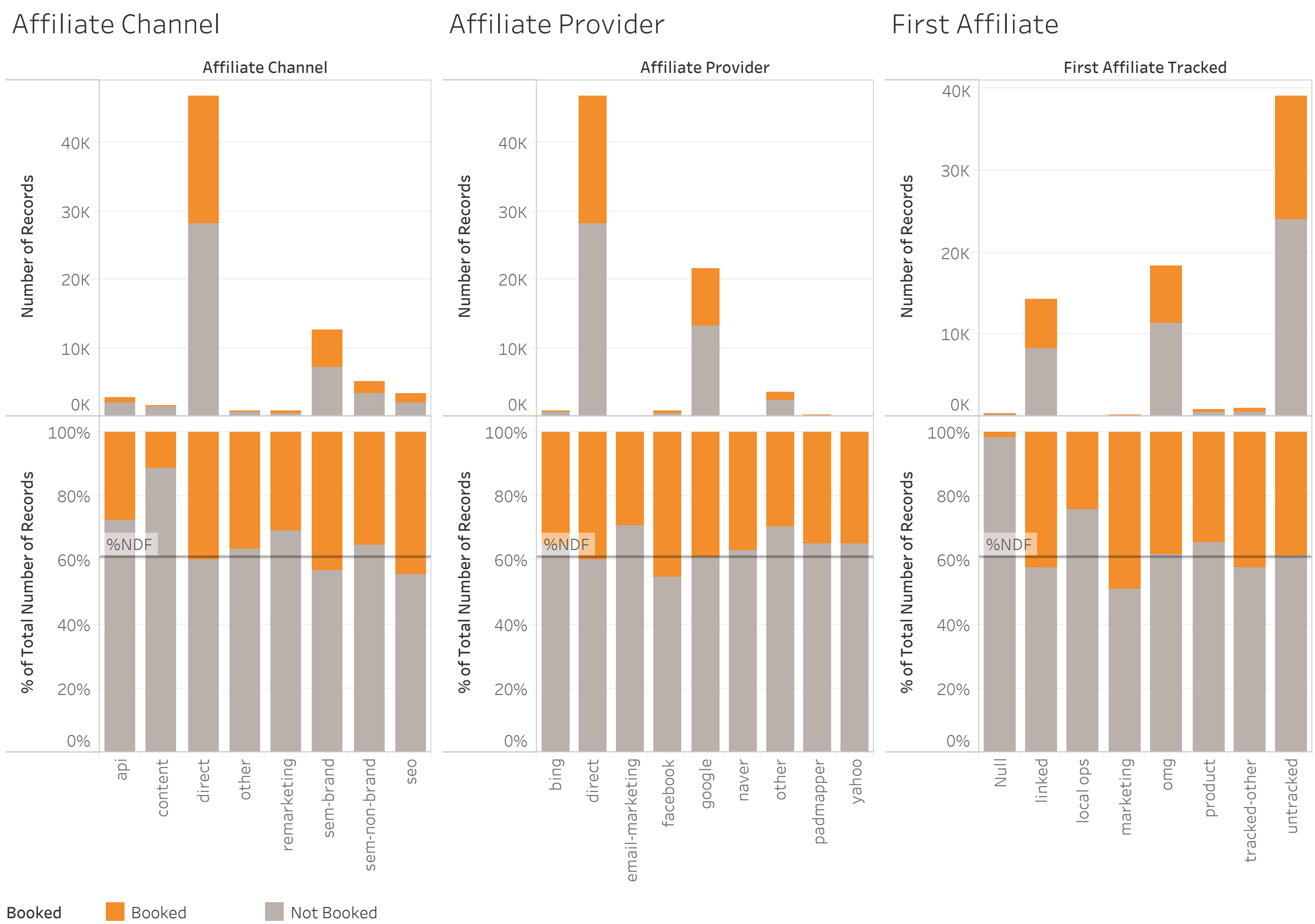
First Browser



Booked Booked Not Booked

Which marketing source was the most effective in resulting in bookings?

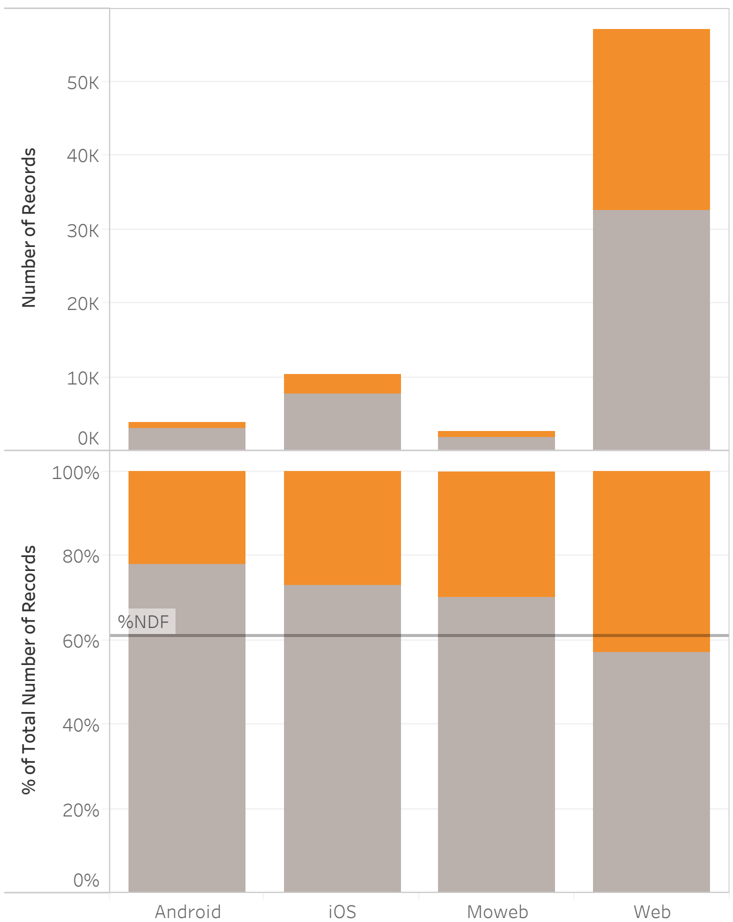
When the affiliate channel is content, or when first affiliate tracked is null, the booking ratio tend to be low.



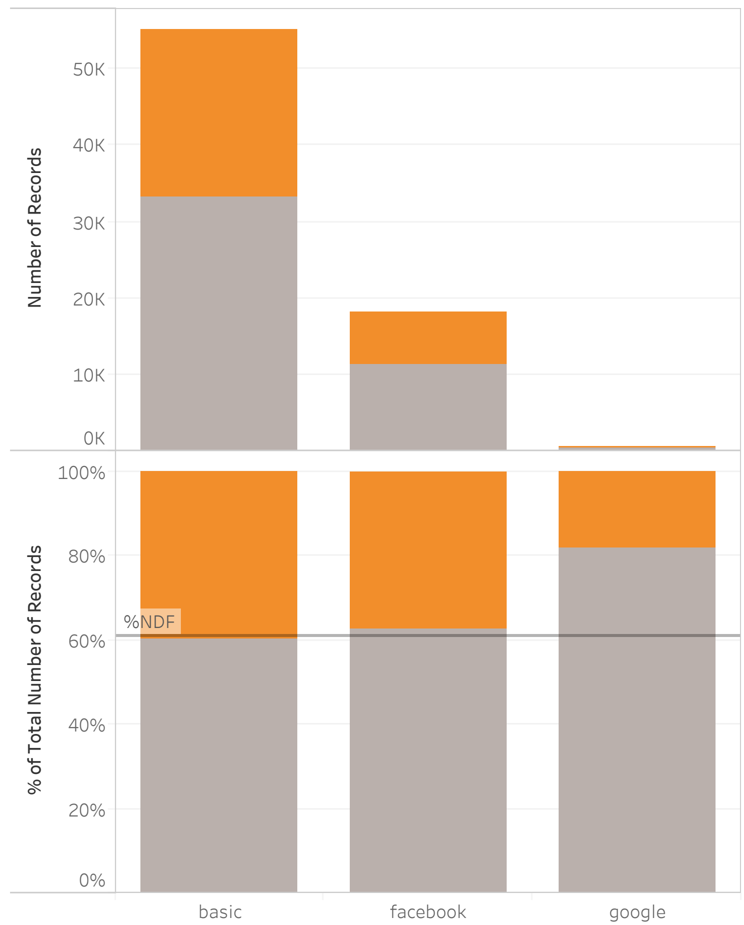
Does Signup App and Signup Method show trend in booking ratios?

People who sign up through web has higher booking ratio. The different signup methods have similar booking ratios.

Signup App



Signup Method

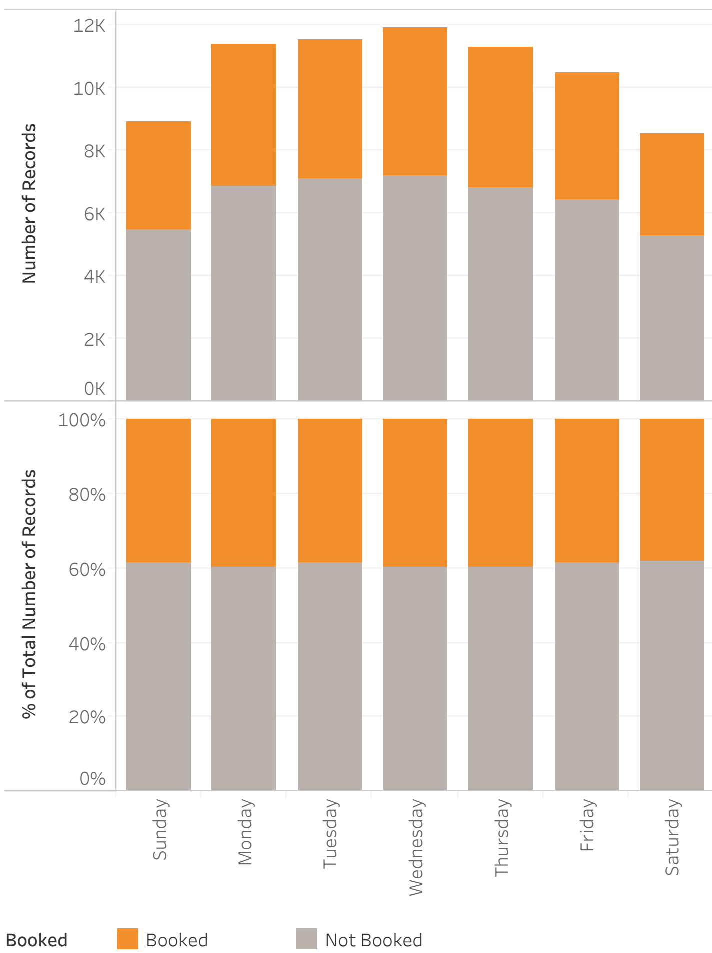


Booked Booked Not Booked

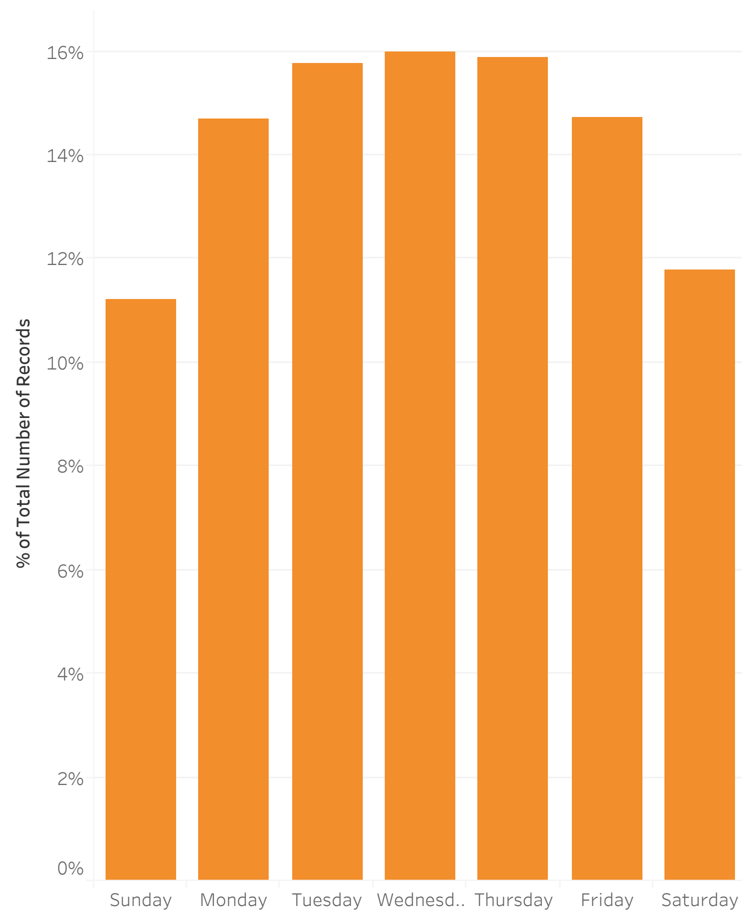
Does day of week affect booking?

Wednesday is the most popular day for account sign up and booking. However, the ratio of booking does not differ across the days of week.

Weekday of Account Creation



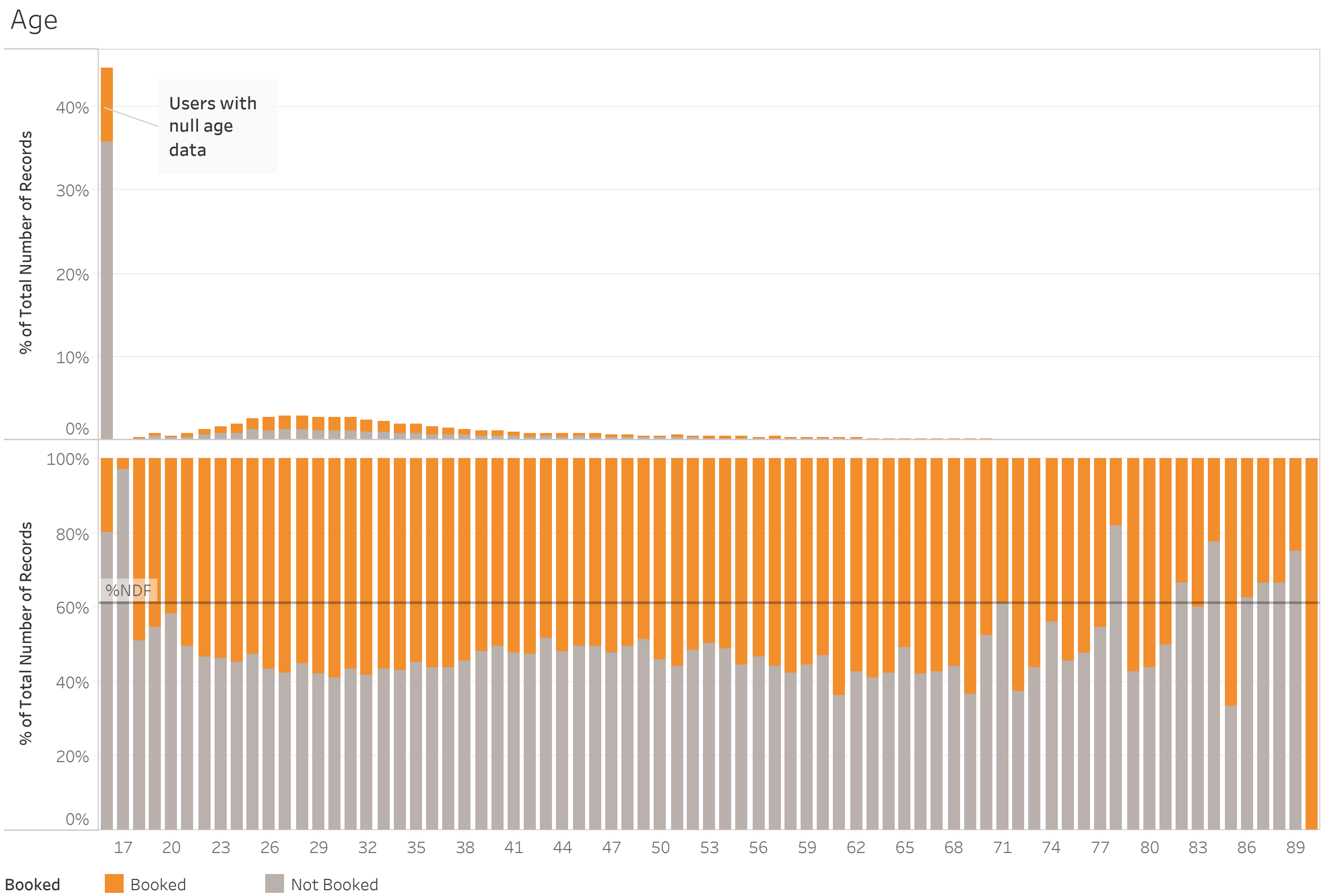
Weekday of First Booking



Data Exploration: Gender | Data Exploration: Device | Data Exploration: Affiliate (marketing channel) | Data Exploration: Sign-up Method | Data Exploration: Day of Week | Data Exploration: Age | Data Exploration: Sessions

Does age affect booking?

People who did not enter age information or entered false age information are less likely to book. The age group with the most users is around 27.

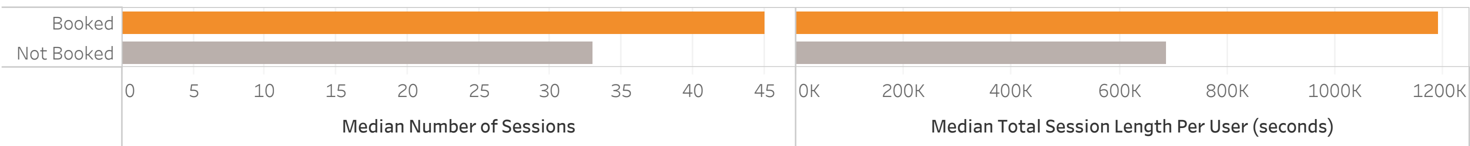


Data Exploration: Gender | Data Exploration: Device | Data Exploration: Affiliate (marketing channel) | Data Exploration: Sign-up Method | Data Exploration: Day of Week | Data Exploration: Age | Data Exploration: Sessions

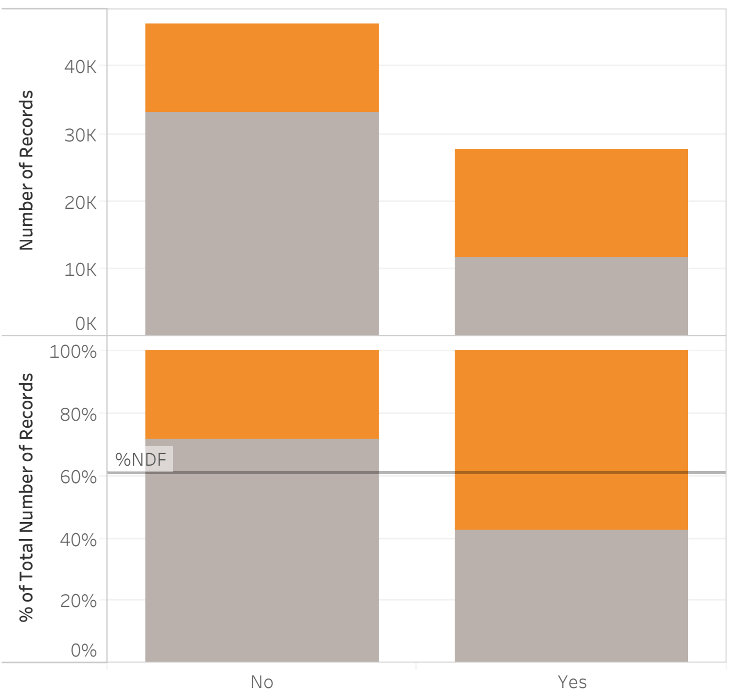
Examples of what we can learn from sessions information?

Users who spent longer time on the site, posted messages, or viewed the cancellation policies are more likely to book.

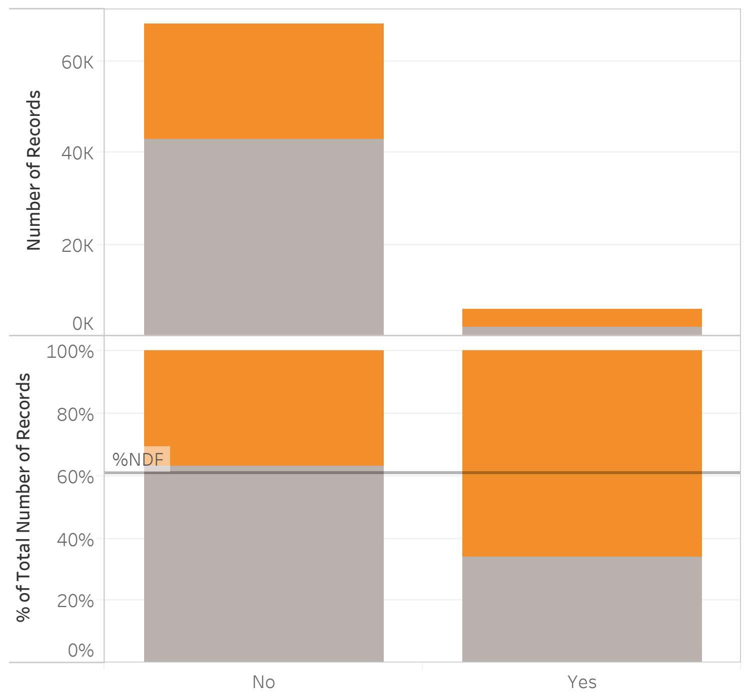
Session Length



Post Message



View Cancellation Policies



Booked Booked Not Booked