A Better Vacation

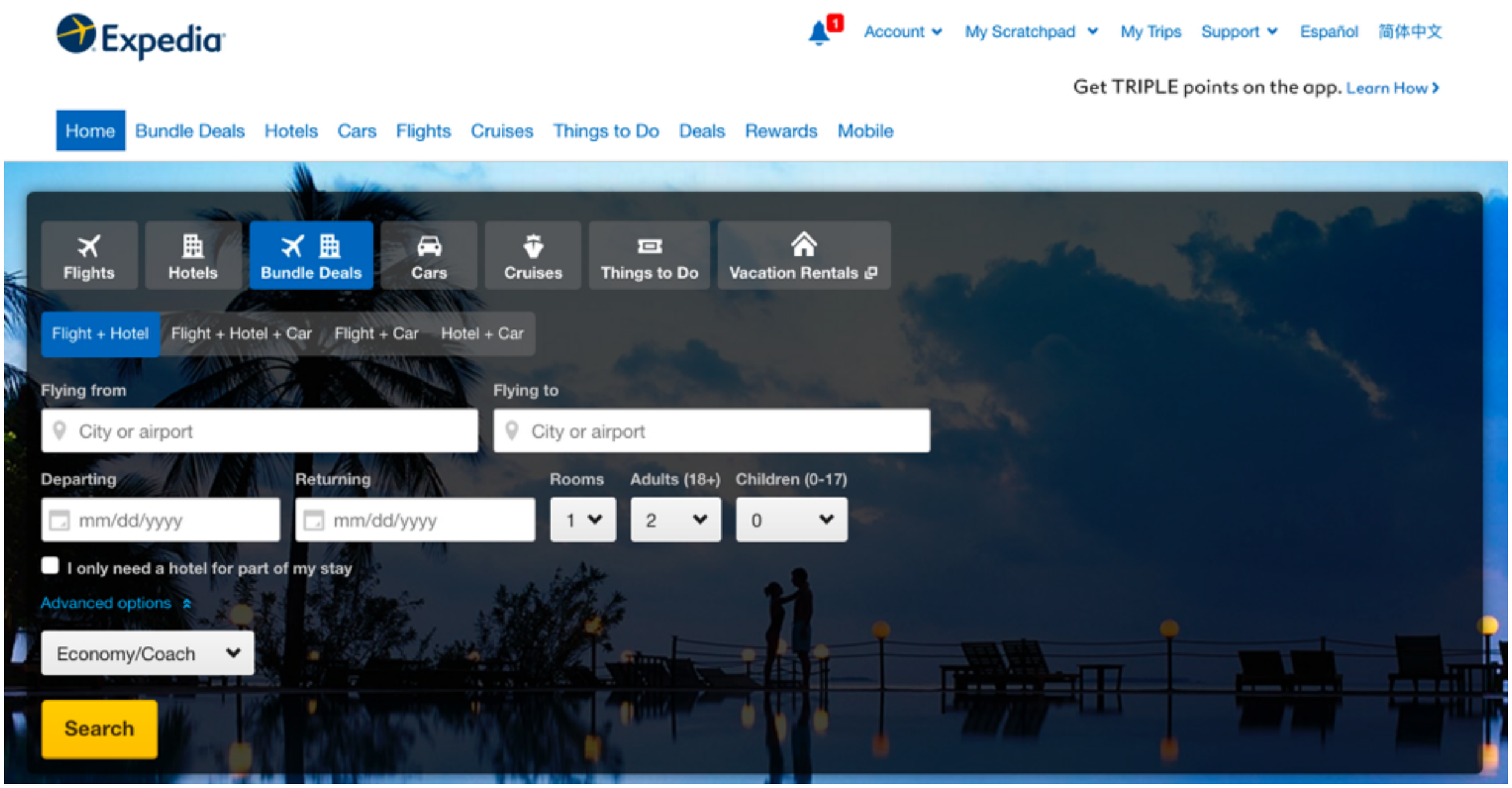
Jared Tom

Expedia®

What is the Problem?

- Expedia wants to provide personalized recommendations for users
- Limited information given
- Goal: Predict the “hotel cluster” for a user

The Data


Expedia
1
Account
My Scratchpad
My Trips
Support
Español
简体中文
Get TRIPLE points on the app. Learn How
Home
Bundle Deals
Hotels
Cars
Flights
Cruises
Things to Do
Deals
Rewards
Mobile
Flights
Hotels
Bundle Deals
Cars
Cruises
Things to Do
Vacation Rentals
Flight + Hotel
Flight + Hotel + Car
Flight + Car
Hotel + Car
Flying from
City or airport
Flying to
City or airport
Departing
mm/dd/yyyy
Returning
mm/dd/yyyy
Rooms
1
Adults (18+)
2
Children (0-17)
0
I only need a hotel for part of my stay
Advanced options
Economy/Coach
Search

The Data

- From Kaggle
- Anonymized
  - Train (37,670,293 / 24)
  - Test (2,528,243 / 22)
- Target: 100 categories of “hotel cluster”
- For ease of use I chose to use 30,000 observations from the Train set
- https://www.kaggle.com/c/expedia-hotel-recommendations/data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 37670293 entries, 0 to 37670292
Data columns (total 24 columns):
date_time                    object
site_name                    int64
posa_continent               int64
user_location_country        int64
user_location_region         int64
user_location_city           int64
orig_destination_distance    float64
user_id                      int64
is_mobile                    int64
is_package                   int64
channel                      int64
srch_ci                      object
srch_co                      object
srch_adults_cnt              int64
srch_children_cnt            int64
srch_rm_cnt                  int64
srch_destination_id          int64
srch_destination_type_id     int64
is_booking                   int64
cnt                          int64
hotel_continent              int64
hotel_country                int64
hotel_market                 int64
hotel_cluster                int64
dtypes: float64(1), int64(20), object(3)
memory usage: 6.7+ GB
```

Predictors

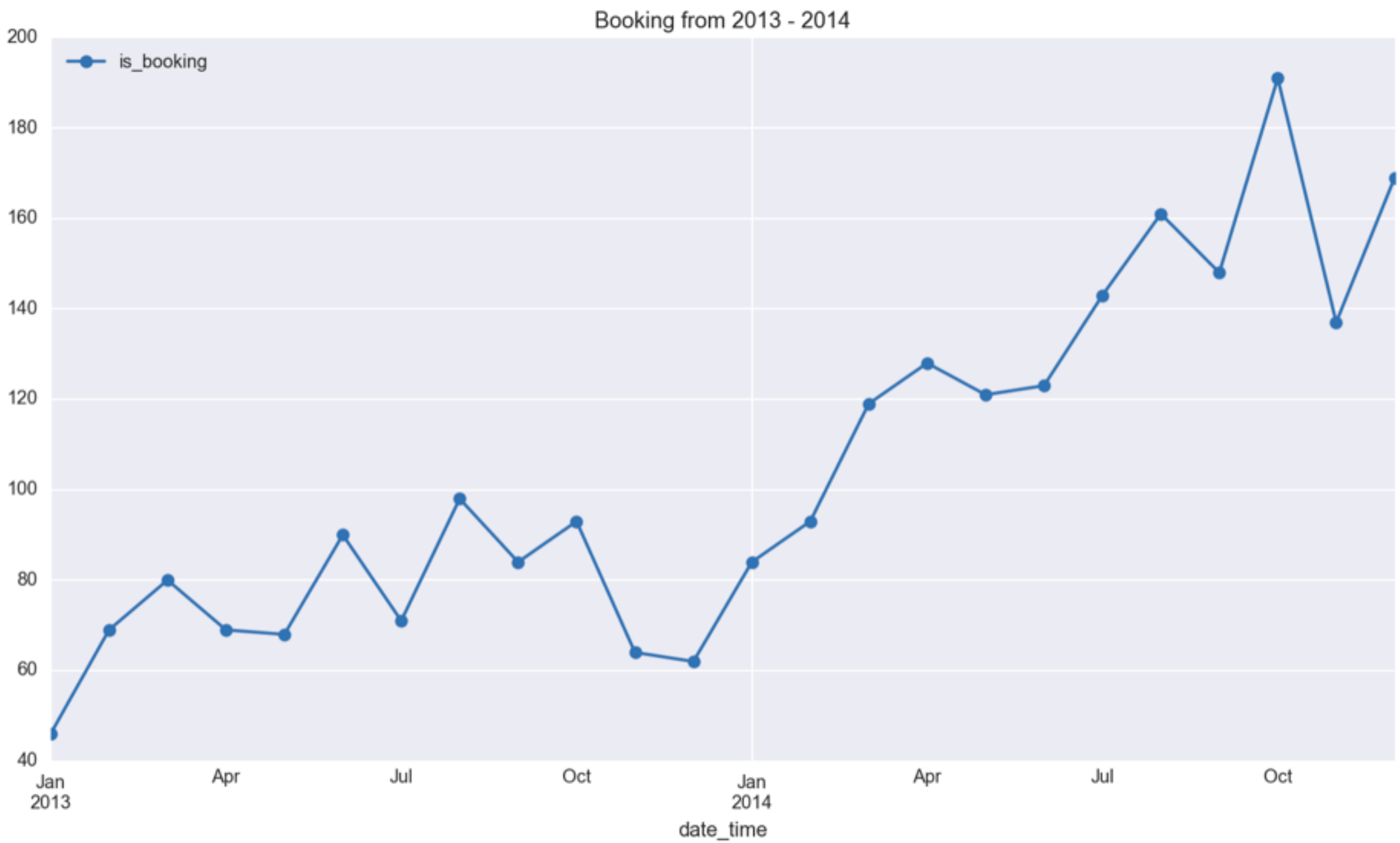
Column name	Description
date_time	Timestamp
site_name	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ...)
posa_continent	ID of continent associated with site_name
user_location_country	The ID of the country the customer is located
user_location_region	The ID of the region the customer is located
user_location_city	The ID of the city the customer is located
orig_destination_distance	Physical distance between a hotel and a customer at the time of search. A null means the distance could not be calculated
user_id	ID of user
is_mobile	1 when a user connected from a mobile device, 0 otherwise
is_package	1 if the click/booking was generated as a part of a package (i.e. combined with a flight), 0 otherwise
channel	ID of a marketing channel
srch_ci	Checkin date
srch_co	Checkout date
srch_adults_cnt	The number of adults specified in the hotel room
srch_children_cnt	The number of (extra occupancy) children specified in the hotel room
srch_rm_cnt	The number of hotel rooms specified in the search
srch_destination_id	ID of the destination where the hotel search was performed
srch_destination_type_id	Type of destination
hotel_continent	Hotel continent
hotel_country	Hotel country
hotel_market	Hotel market
is_booking	1 if a booking, 0 if a click
cnt	Numer of similar events in the context of the same user session
hotel_cluster	ID of a hotel cluster

Exploration

Total Bookings

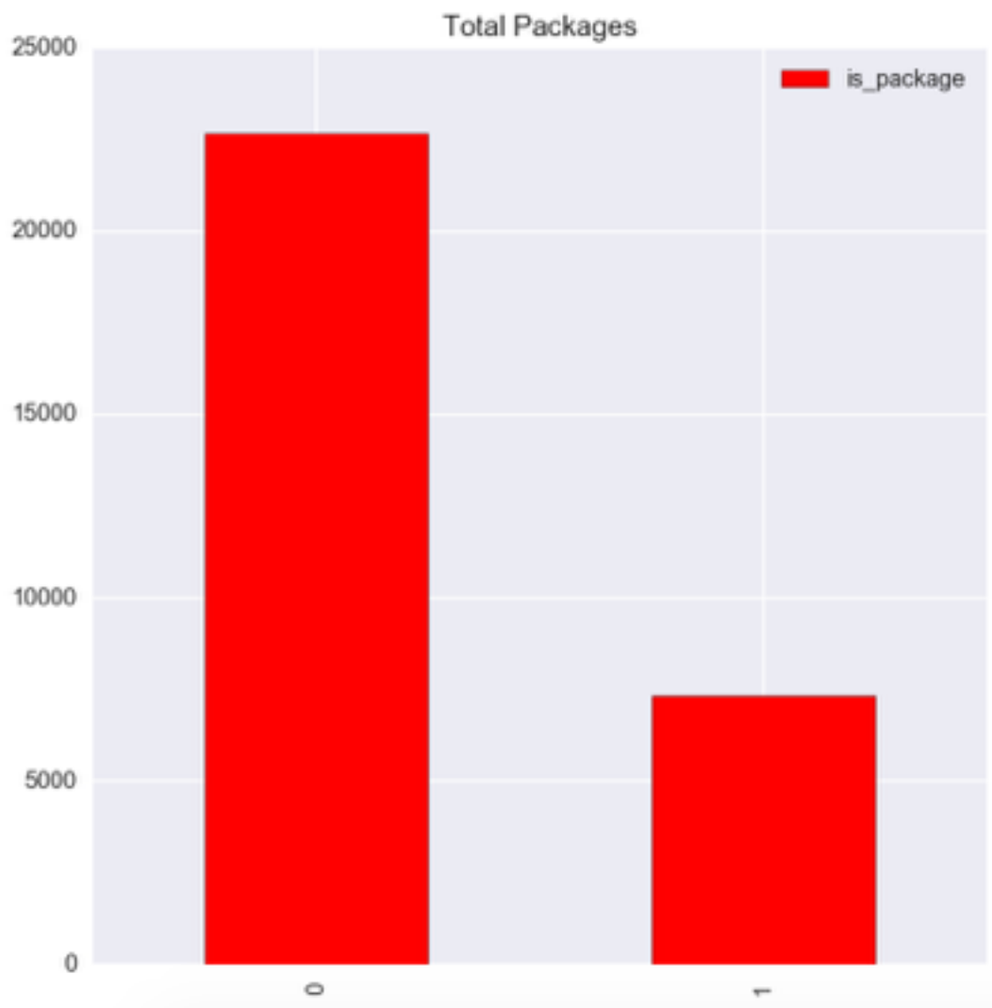
```
0    27489
1     2511
Name: is_booking, dtype: int64

<matplotlib.axes._subplots.AxesSubplot at 0x2e08c0d50>
```

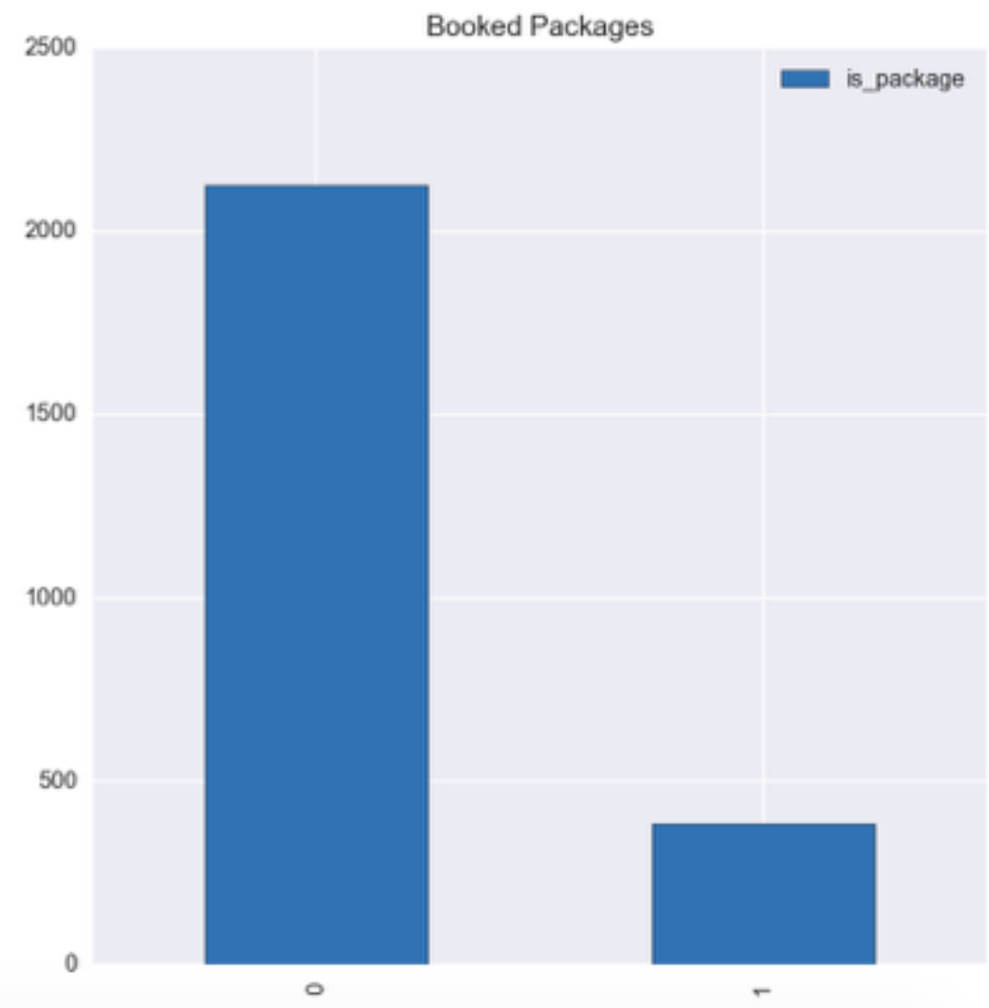
Booking Frequency



Packages

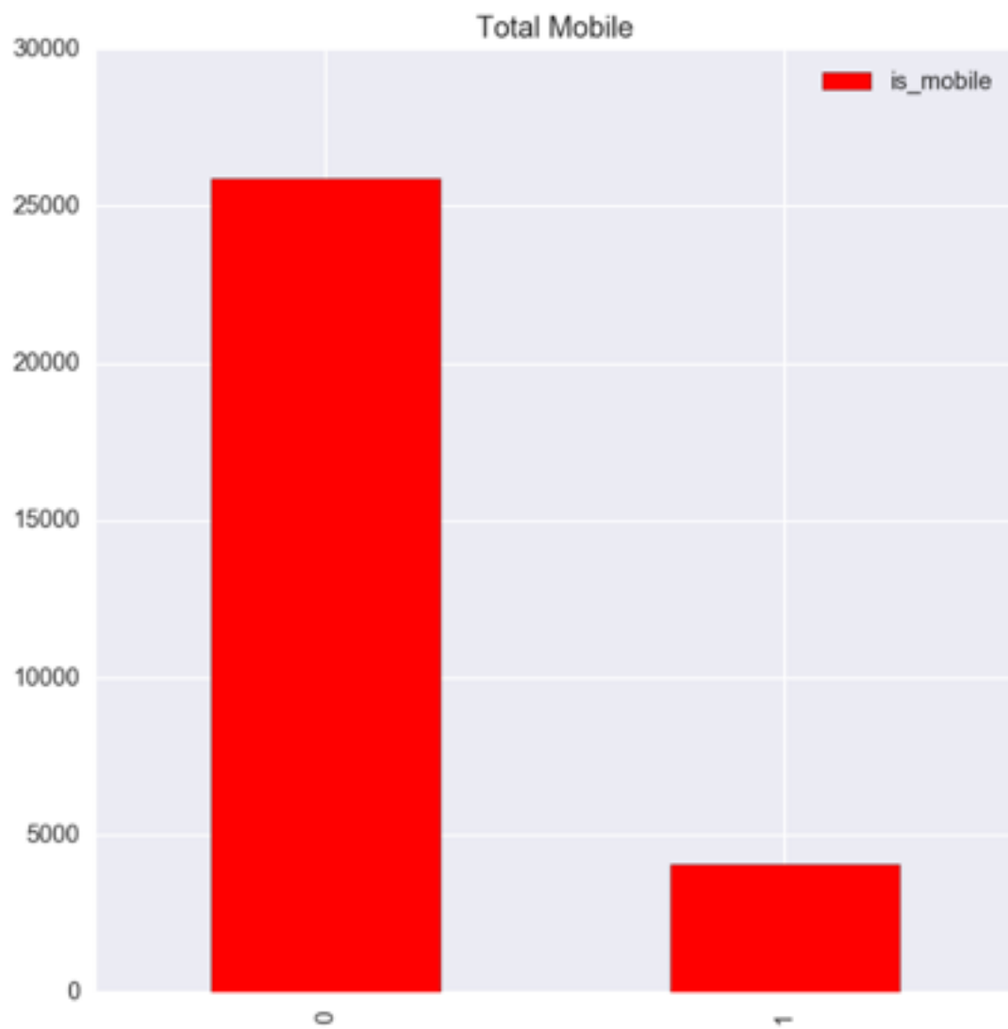
```
0    22658
1     7342
Name: is_package, dtype: int64
----------------------------
0    2126
1     385
Name: is_package, dtype: int64

<matplotlib.axes._subplots.AxesSubplot at 0x313731090>
```

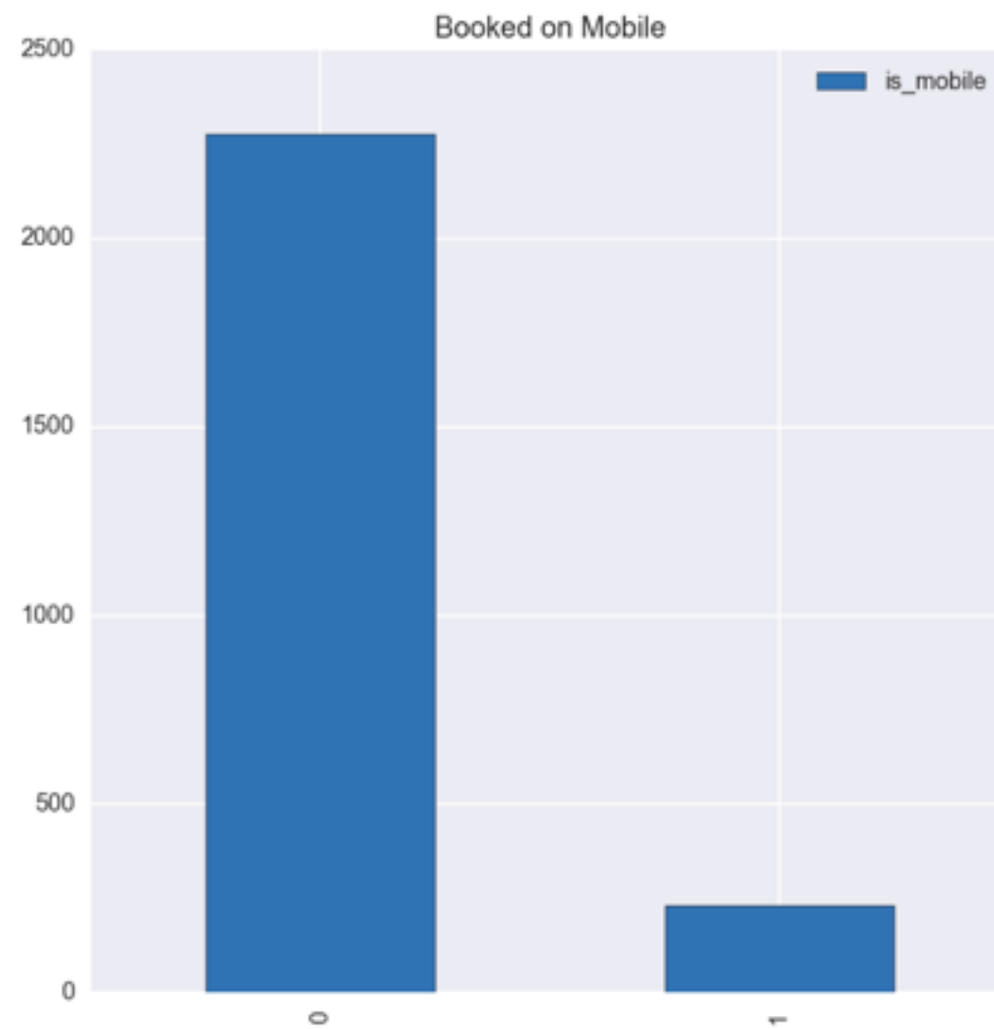




Mobile

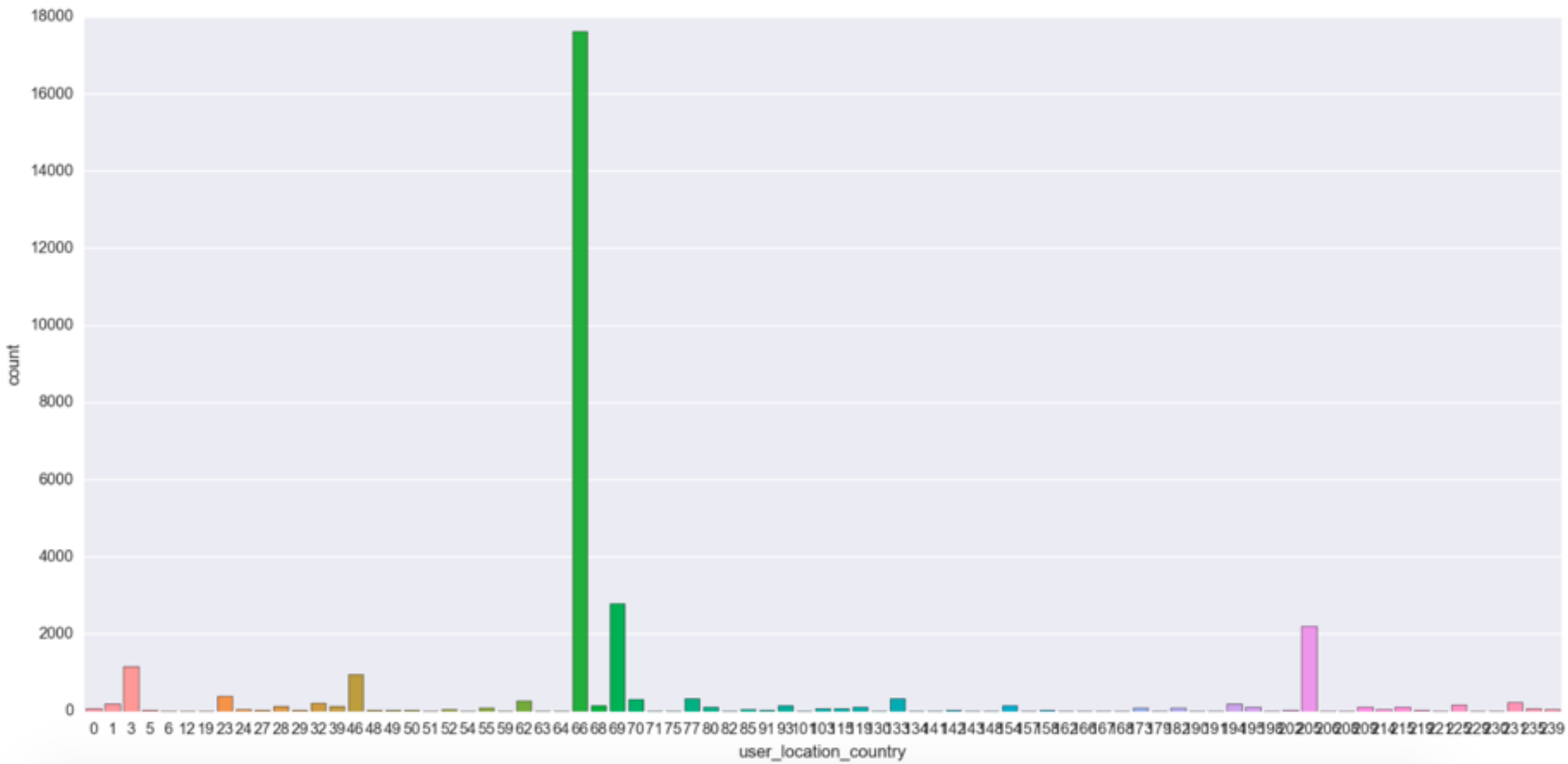
```
0    25919
1     4081
Name: is_mobile, dtype: int64
----------------------------
0    2278
1     233
Name: is_mobile, dtype: int64

<matplotlib.axes._subplots.AxesSubplot at 0x2b92d2390>
```





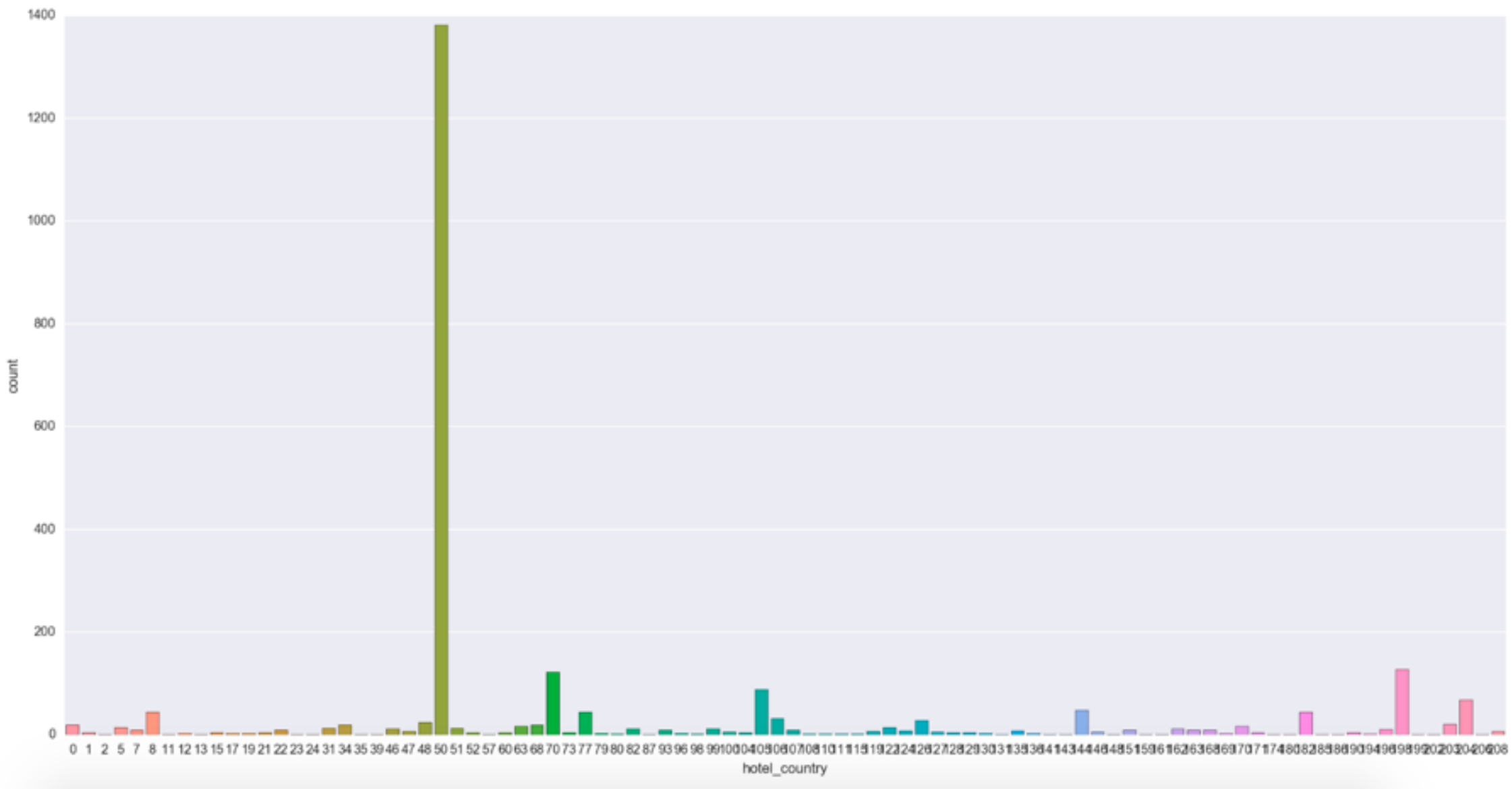
Country with the Most Users

```
66     17632
69      2797
205     2201
3       1166
46       951
Name: user_location_country, dtype: int64

<matplotlib.axes._subplots.AxesSubplot at 0x2f474dc10>
```



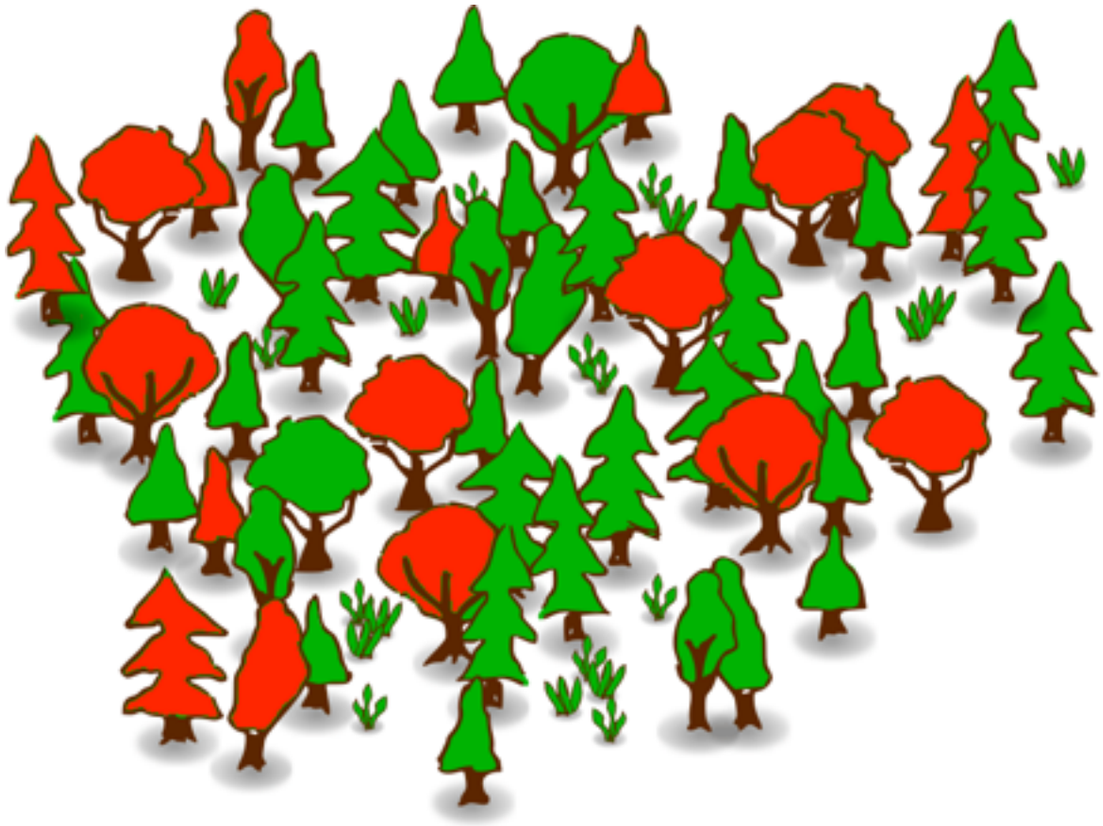
Most Traveled Country

```
50     1382
198     127
70      122
105      88
204      68
Name: hotel_country, dtype: int64

<matplotlib.axes._subplots.AxesSubplot at 0x12a84b910>
```



Models

- Chosen Models:
    - Decision Tree: quick, possible overfit
    - Random Forest: multiple trees
    - KNeighbors Classifier: similar neighbors, not the most accurate



Models

Decision Tree

	feature	importance
17	hotel_market	0.209313
15	hotel_continent	0.111541
5	user_id	0.089478
18	day	0.08704
4	user_location_city	0.070052
19	month	0.067275
3	user_location_region	0.065856
16	hotel_country	0.062527
11	srch_destination_id	0.059552
14	cnt	0.05198
12	srch_destination_type_id	0.025395
8	srch_adults_cnt	0.020946
9	srch_children_cnt	0.016735
2	user_location_country	0.016288
0	site_name	0.013351
7	is_package	0.010659
13	is_booking	0.007102
6	is_mobile	0.00555
1	posa_continent	0.005535
10	srch_rm_cnt	0.003825

Random Forest

	feature	importance
17	hotel_market	0.204193
15	hotel_continent	0.107242
18	day	0.094098
5	user_id	0.086837
4	user_location_city	0.075392
19	month	0.068232
11	srch_destination_id	0.065803
3	user_location_region	0.063853
16	hotel_country	0.062648
14	cnt	0.045559
12	srch_destination_type_id	0.025082
8	srch_adults_cnt	0.020908
2	user_location_country	0.016191
9	srch_children_cnt	0.014849
0	site_name	0.014188
7	is_package	0.011737
6	is_mobile	0.006747
13	is_booking	0.006664
1	posa_continent	0.005756
10	srch_rm_cnt	0.004021

Models

- There was not a strong separation between the models.
  - Decision Tree: 27%
  - Random Forest: **29%**
  - KNeighbors Classifier: 28%

Final Results

- The data did not have a significant amount of variance
- All three models used were significantly better than the baseline, but gave between 27 - 29%
- Need more info to give personalized recommendations

Call to Action

- Give a few more filters in the initial search
    - price, rating
- Build a recommender system so people do not have to search through multiple pages for their “best” hotel and flight
    - recommendation system might translate to higher conversion rate

Next Steps

- Move to AWS or Spark to handle entire dataset
- Explore specific countries more in depth
- Use SVC for a future model
- Collaborative Filtering
- Use location as one training set and take out location in another set