

# Airbnb New User Bookings Project - Milestone Report

## Objective:

Airbnb's business covers 34,000+ cities across 190+ countries. Thus, being able to accurately identify where new users are heading to is important as it allows Airbnb to recommend rentals that suit customers' needs and as a result, decrease the average time to first booking and improve the site's booking rate overall.

Research Question: Predict which country new users will book their first trip based on users' demographic data, web session records, and some summary statistics of different countries.

## Data:

The datasets for this project are available on Kaggle (<https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings>)

These datasets are:

- train\_users.csv: training data. The file includes information related to Airbnb accounts such as when the user signed up/ made his/her first booking, sign-up flow, language preference, etc.
- test\_users.csv: testing data. It contains the same fields as train\_users.csv's except that test\_users.csv does not include the outcome variable which is the destination country for the first booking.
- age\_gender\_bkts.csv: It includes different countries' age/ gender splits.
- countries.csv: It includes geographic information of different countries.
- sessions.csv: It contains users' web session log.

## Data Wrangling:

For train\_users\_2 & test\_users datasets:

- Parse datetime variables 'date\_account\_created', 'date\_first\_booking' and 'timestamp\_first\_active' to pandas datetime objects.
- Replace 'unknown-' with numpy NaN.
- 0.01% accounts were created after the first trip was booked ('date\_account\_created' > 'date\_first\_booking'). According to Airbnb's terms & conditions, users need to register for an Airbnb account first before booking and so I replace the 'date\_account\_created' with 'date\_first\_booking' for these observations.
- 1% account holders are > 100 years old while 0.068% are < 18 years old. Since Airbnb requires users to be at least 18 years old in order to use the site (<https://www.airbnb.com/terms>), I replace all these < 18 years old with 18. I also replace those > 100 years old with 100.
- Create variables that capture the year, month, day, week and day of week of 'date\_account\_created', 'date\_first\_active'.
- We have more than 40% NaNs for age & gender. Use machine learning to predict these missing values from other variables because these two variables are need to join the age\_gender\_bkts dataframe. Create a binary indicate to record if their original values

are NaNs since the fact that these values are missing might tell us something about the data.

- Also use machine learning to fill out missing values for language.
- Drop 'date\_first\_booking' since this feature is not available in the test set (all NaN).
- Do one-hot encoding for categorical variables. For each categorical variable, include an indicator that records if the value of the observation is missing.

For age\_gender\_bkts:

- Transform the dataframe from long to wide format for merging with train & test data.

For countries datasets:

- Variable 'destination\_language' is renamed 'destination\_language'.
- Values of 'destination\_language' is mapped to the 'language' variables in train & test sets.

For session dataset:

- Replace '-unknown-' with numpy NaN.
- Observations with missing user\_ids are dropped since it is the key variables for merging with the train/ test data.
- 1.7% sessions with session time > 200,000 seconds (55.56 hours). Leave them as is and will probably try log transformation during modeling.
- Drop observations with missing session time as they probably are not legit sessions
- Create a dataframe (df\_session2) that records for each user:
  - Total number of actions taken
  - Total number of unique Actions, ActionTypes, Action\_Details, Devices
  - Sum, mean, min, max, median, s.d., skewness, kurtosis of seconds\_elapsed
  - Number and proportion (sum & mean) of each type of action taken, e.g. 'lookup', 'search\_results'

Merging datasets:

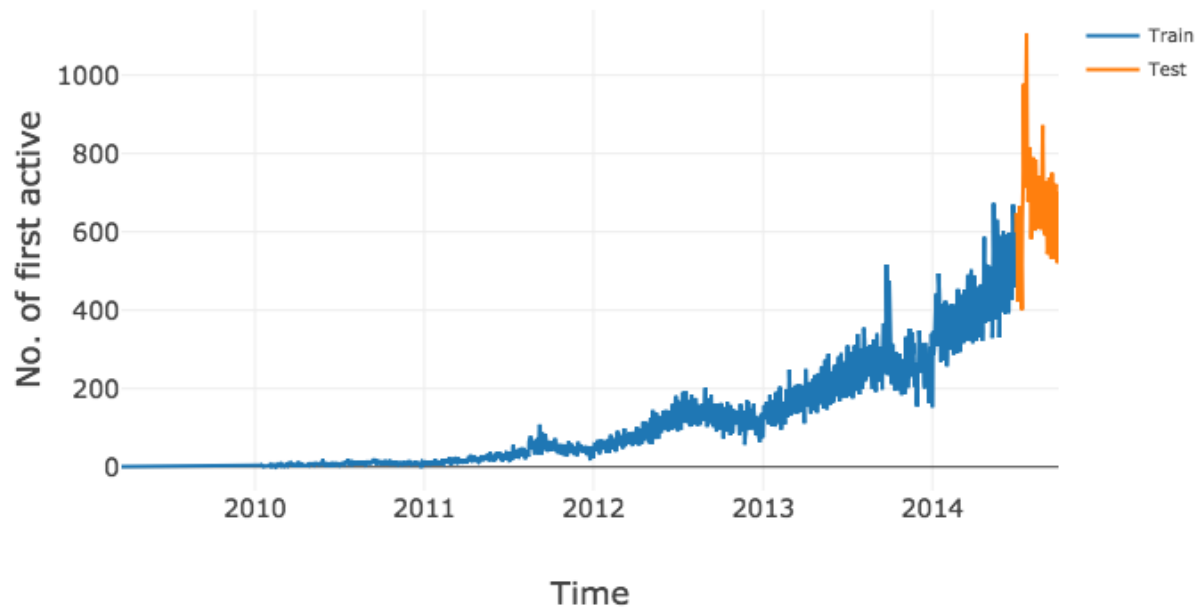
- Merge train\_users\_2 & test\_users datasets with age\_gender\_bkts using age and gender.
- For the merged datasets, create a variable that records if the user's preferred language matches the language of each country using the information from the countries dataframe
- Join the merged datasets with the session data (df\_session2) using user\_id

After merging the datasets:

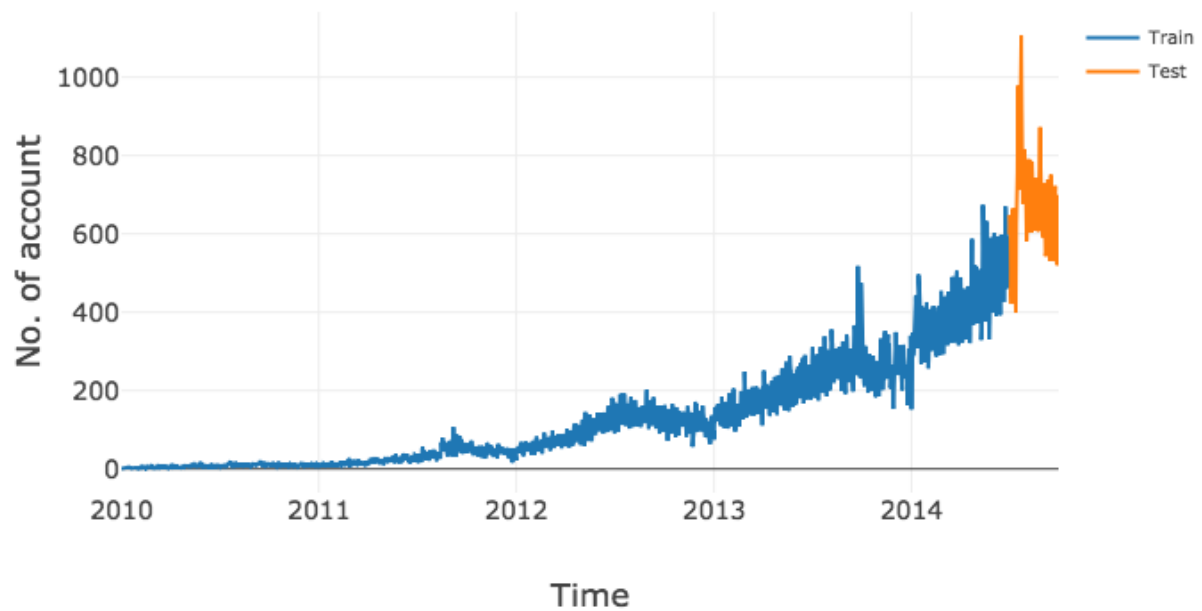
- Some user\_id appear in training & test datasets but not in session data and so we have lots of NaNs for variables related to sessions after merging the dataset. Replace these NaNs with 0.

## Exploratory Analysis:

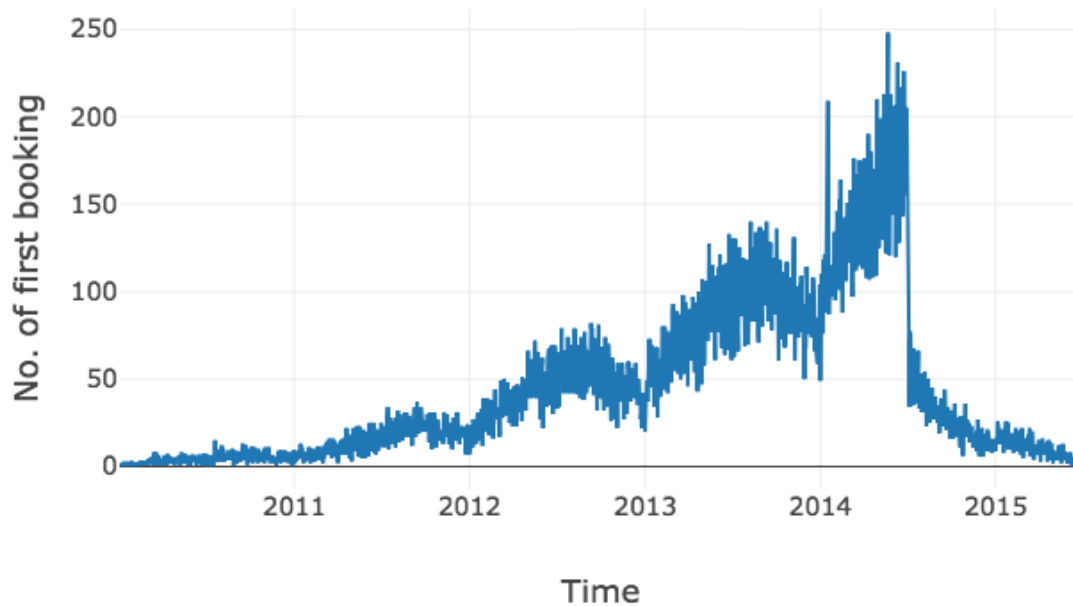
### No. of First Active over Time



### No. of Account Created over Time



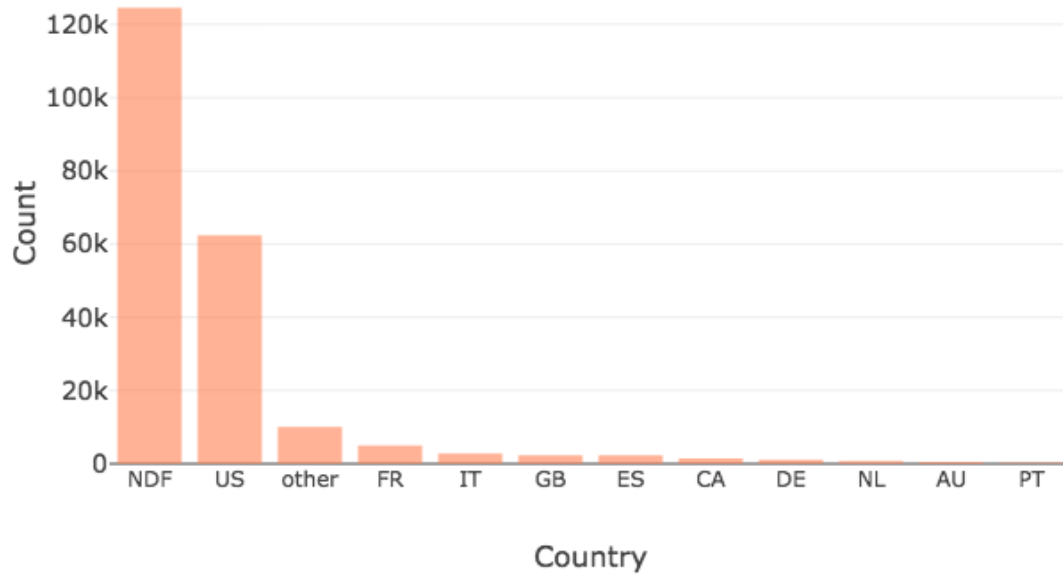
## No. of First Booking over Time



- Training set contains users with first active date and account creation date before 2014-07-01 while test set contains those with first active date and account creation date on or after 2014-07-01.
- Seems like in mid July 2014, there were an influx of new users trying and signing up for Airbnb without actually booking their first trip.

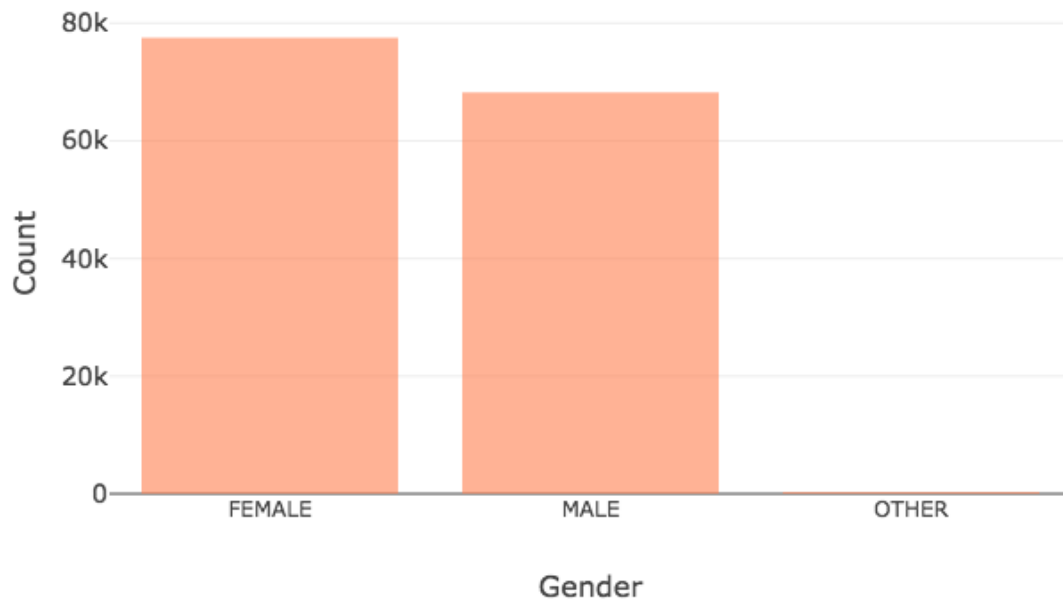
Let's look into these Airbnb users in terms of who they are and where they booked their first trip.

## Distribution of Destination

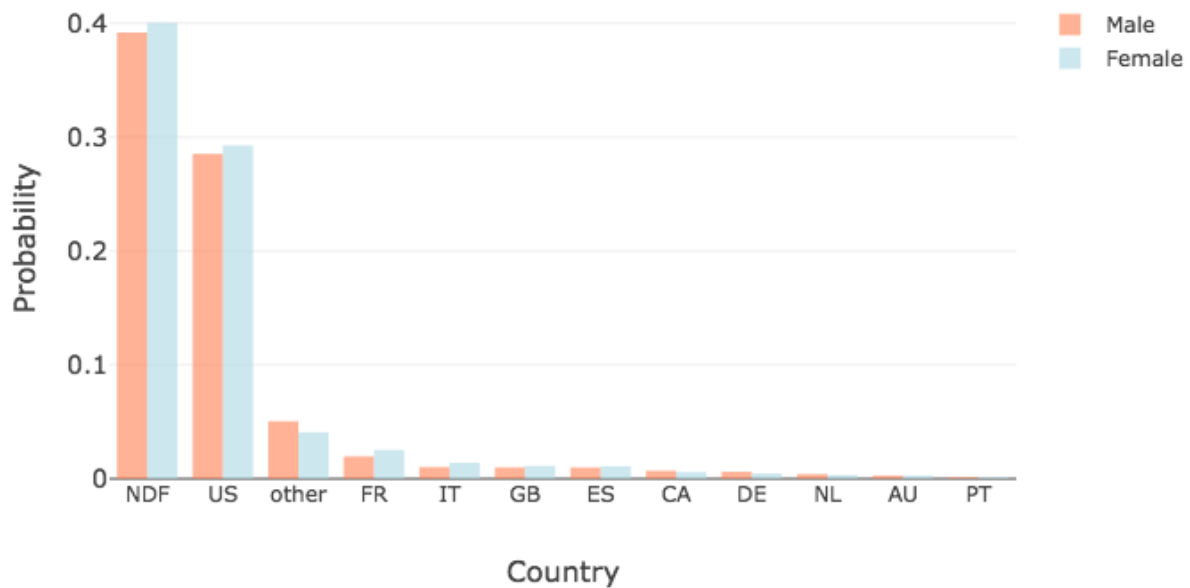


- Most users did not end up booking a trip; for those who ended up booking, most of them booked a trip to the U.S.

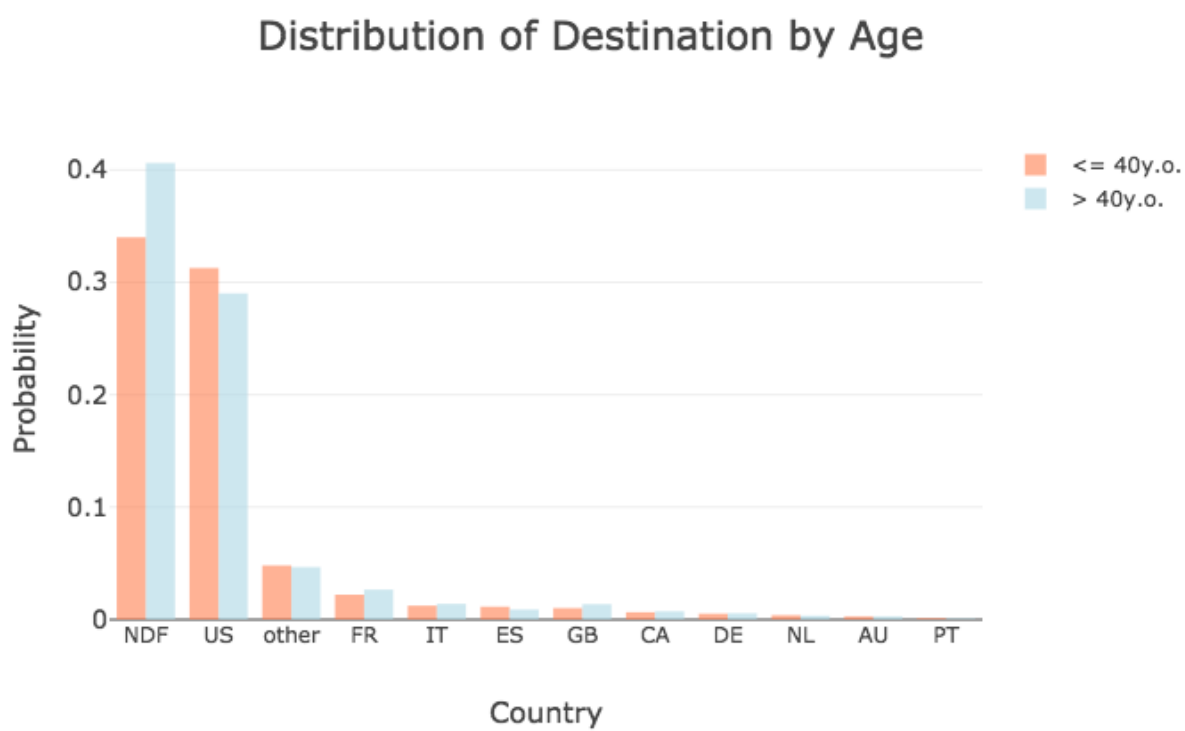
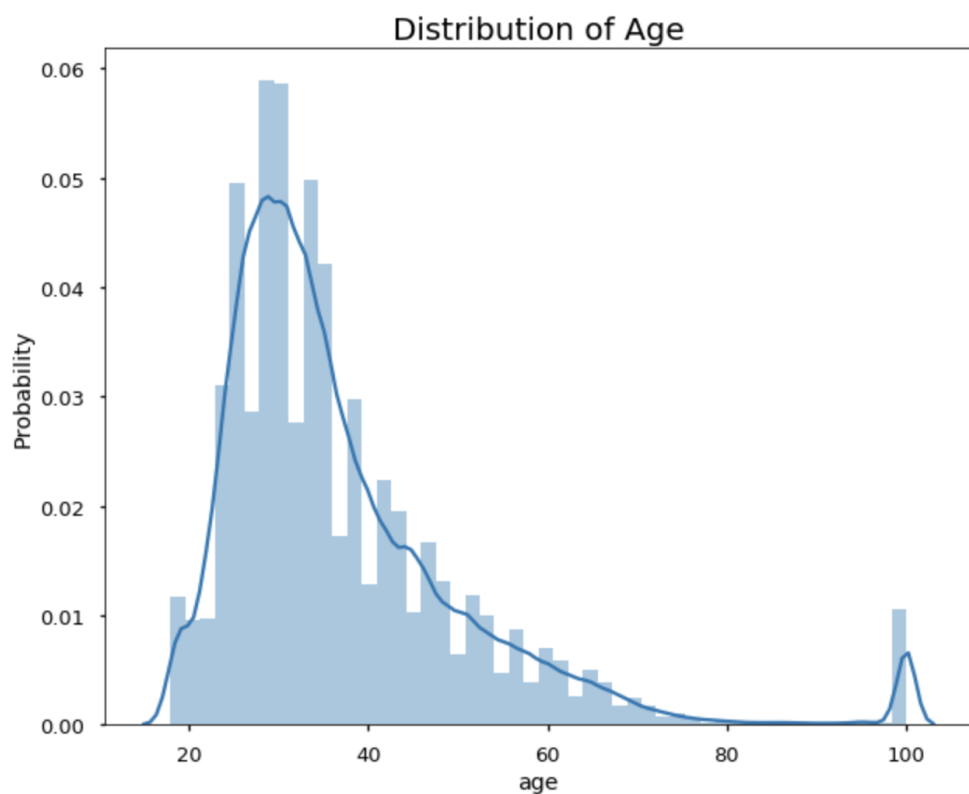
## Distribution of Gender



## Distribution of Destination by Gender



- We have slightly more female users than male. However, male and female users did not seem to show different preference when picking their first destination.



- The majority of Airbnb users are under 40. Younger users ( $\leq 40$  years old) had a higher probability of booking a trip (i.e. lower NDF) than those who are  $> 40$  years old. A one-tail z test is conducted to test for it.

Null hypothesis: Probability of younger users ( $\leq 40$  years old) booking a trip - Probability of older users ( $> 40$  years old) booking a trip = 0

Alternative hypothesis: Probability of younger users ( $\leq 40$  years old) booking a trip - Probability of older users ( $> 40$  years old) booking a trip  $> 0$

The p-value of our test is  $7.20e-68$ . Since the p-value is very close to 0, we reject the null hypothesis and conclude that younger users had a higher probability of booking a trip.
- Younger users were also more likely to pick U.S. as the destination of their first trip. Again, a one-tail z test is conducted to test for the below null & alternative hypothesis

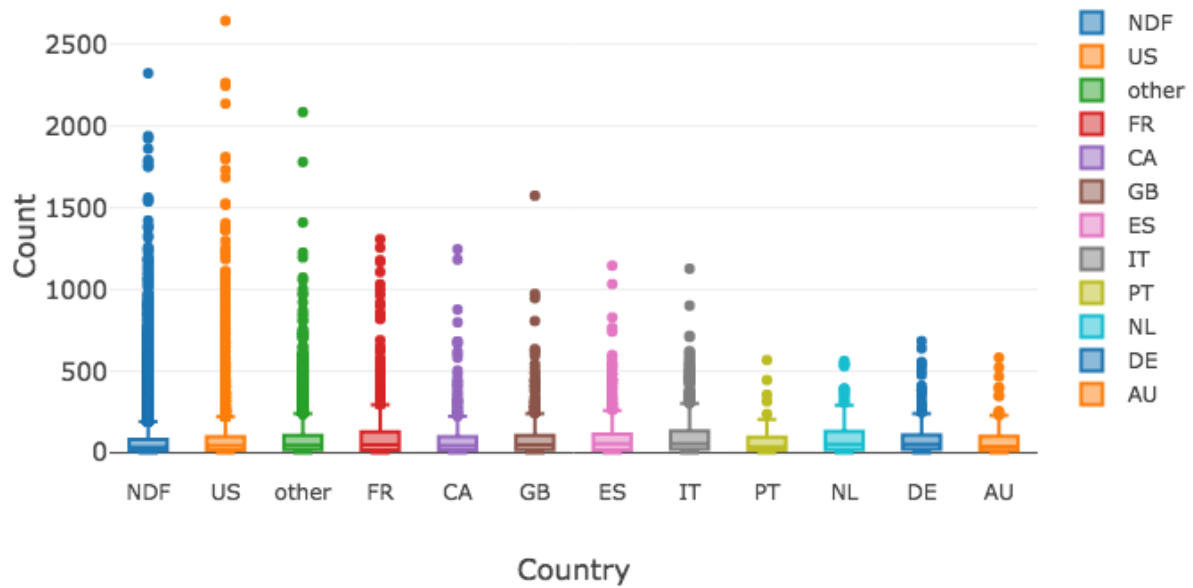
Null hypothesis: Probability of younger users ( $\leq 40$  years old) booking a trip to US - Probability of older users ( $> 40$  years old) booking a trip to US = 0

Alternative hypothesis: Probability of younger users ( $\leq 40$  years old) booking a trip to US - Probability of older users ( $> 40$  years old) booking a trip to US  $> 0$

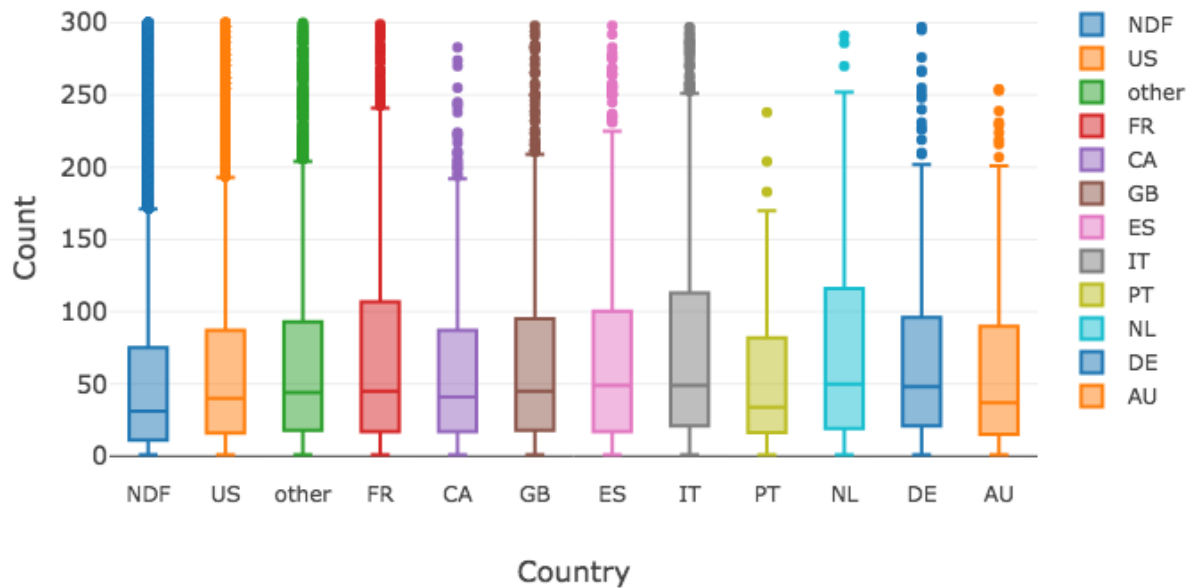
The p-value of our test is  $3.45e-66$ . We reject the null hypothesis because the p-value is very close to 0 and conclude that younger users had a higher probability of booking a trip to the U.S.



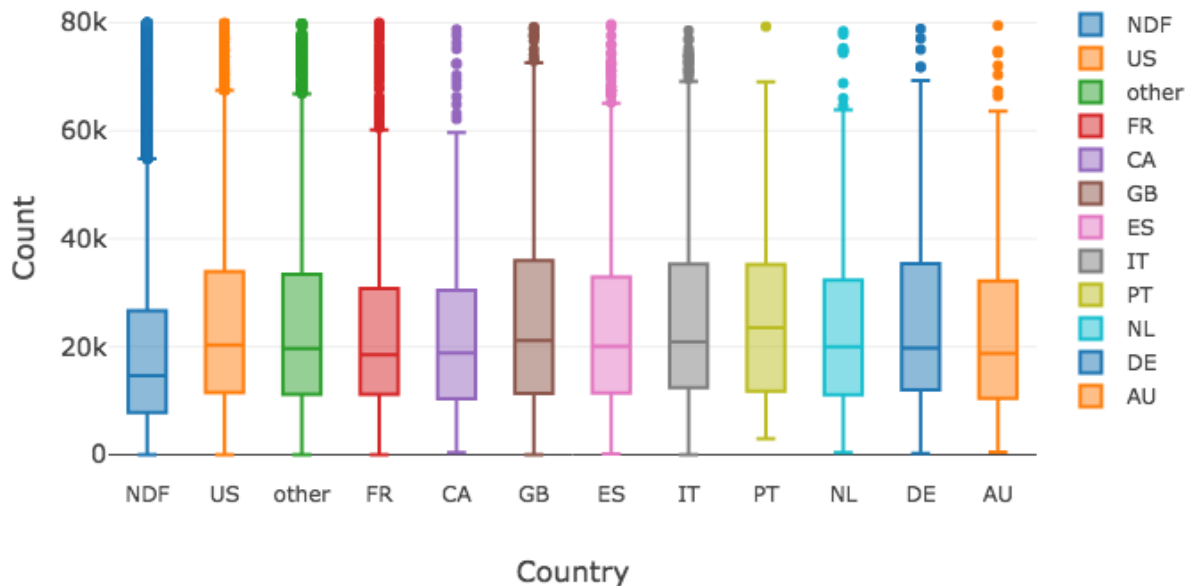
No. of Web Session by Destination



No. of Web Session by Destination ( $\leq 300$  Sessions)



## Web Session Time by Destination (<= 80000 Sec)



- Regarding the number and length of web sessions, seems like those who didn't book a trip had less and shorter sessions.

### Those who booked a trip had more web sessions

- Null hypothesis: Mean number of web sessions among those who booked a trip - Mean number of web sessions among those who did not book a trip = 0
- Alternative hypothesis: Mean number of web sessions among those who booked a trip - Mean number of web sessions among those who did not book a trip > 0
- A one-tail z test is conducted to test for the mean no. of sessions between those who booked a trip vs those who didn't. Z test is used because our sample sizes are larger than 30. It is a one-tailed test because we are interested in whether the one mean is greater than the other.
- The p-value of our test is 5.12e-104. Since the p-value is very close to 0, we reject the null hypothesis and conclude that those who booked a trip visited Airbnb's website/ app more frequently.

### Those who booked a trip had longer web sessions

- Null hypothesis: Mean web session length among those who booked a trip - Mean web session length among those who did not book a trip = 0
- Alternative hypothesis: Mean web session length among those who booked a trip - Mean web session length among those who did not book a trip > 0

- Again, a one-tail z test is conducted to test for the mean session length between those who booked a trip vs those who didn't because our sample sizes are larger than 30 and we are interested in whether the one mean is greater than the other.
- The p-value of a one-tail z test is  $4.68e-34$ . We reject the null hypothesis and conclude that those who booked a trip had longer web sessions.