WILL NEW AIRBNB USERS BOOK THEIR FIRST TRIP?

Brought to you by
Anne T Griffin

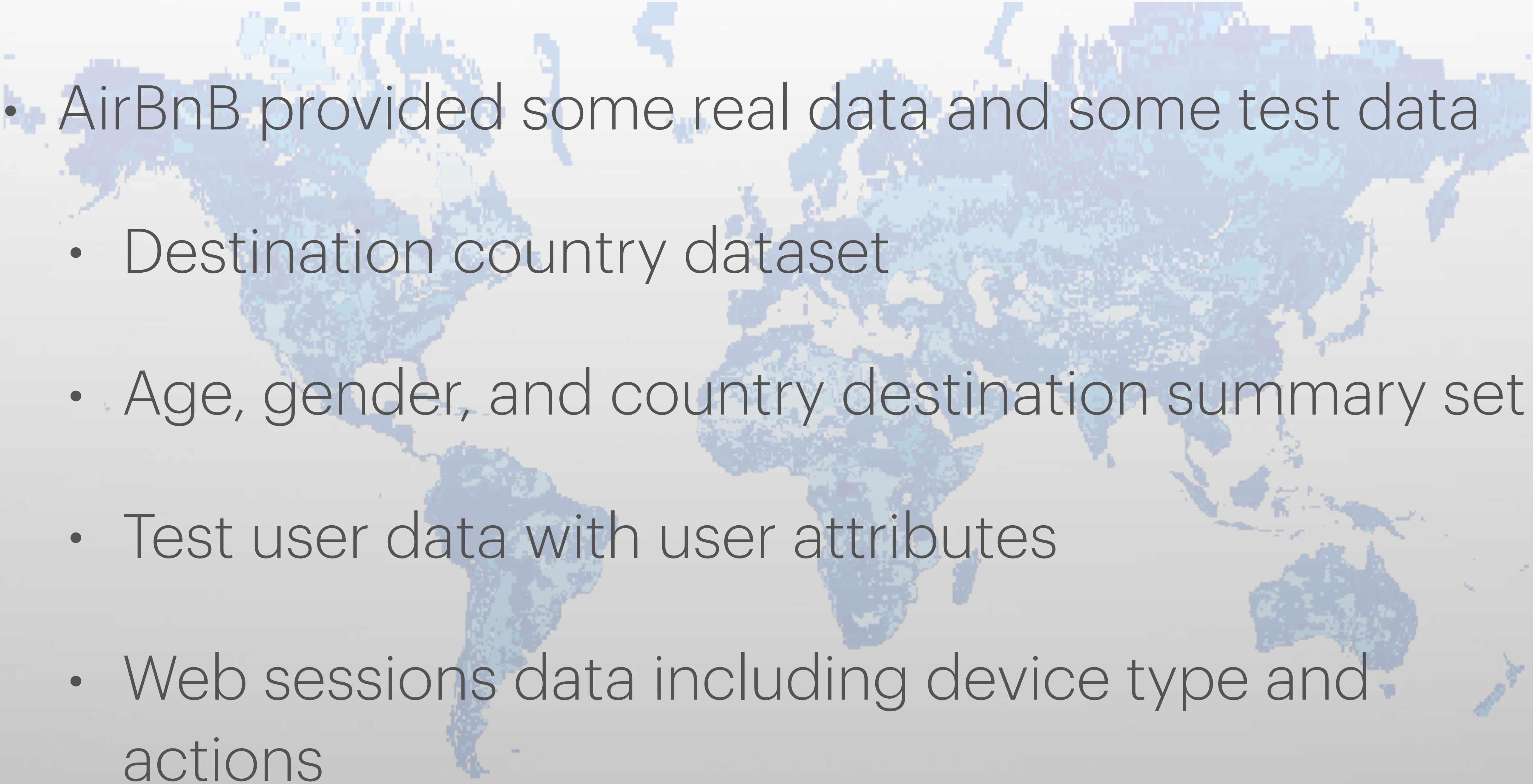
WHY DOES AIRBNB WANT TO KNOW THIS?

GROWTH

WHY DOES AIRBNB WANT TO KNOW THIS?

- Convert new users to active users
  - Personalize content to where they are likely to book
    - Reduce time between sign up and first booking

THE DATA



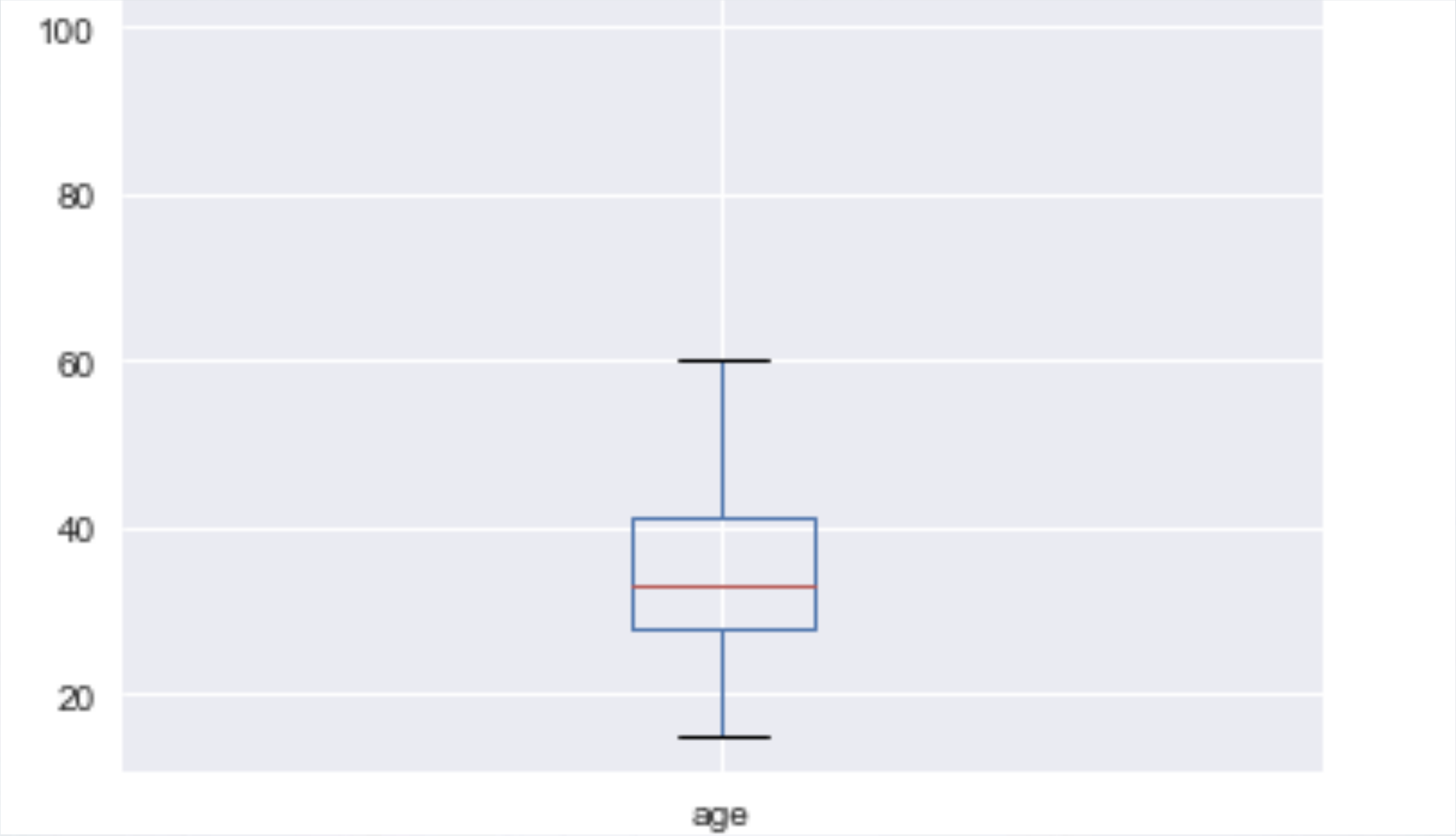
- AirBnB provided some real data and some test data
    - Destination country dataset
    - Age, gender, and country destination summary set
    - Test user data with user attributes
    - Web sessions data including device type and actions

NULL HYPOTHESIS

There is **no correlation** between a user’s age, gender, and other user info and the destination country they will first book.

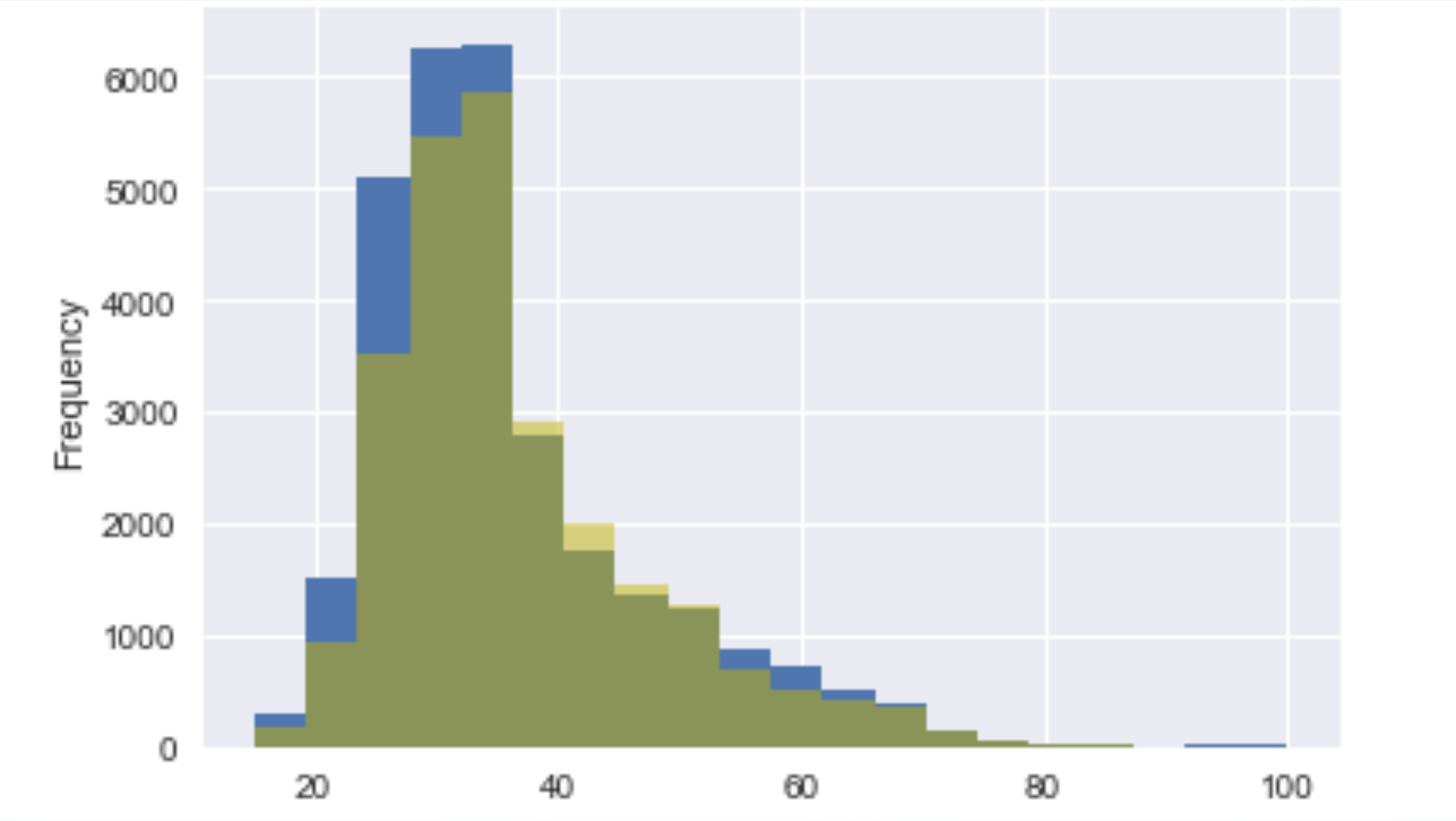
ALTERNATIVE HYPOTHESIS

Using basic info about the user such as age, gender, and other attributes such as device and when and how they signed up, we can predict which country where they are likely to book their first trip.



AGE DISTRIBUTION

Originally, had several outliers around 100. Some values were “2014”, so those needed to be dropped.

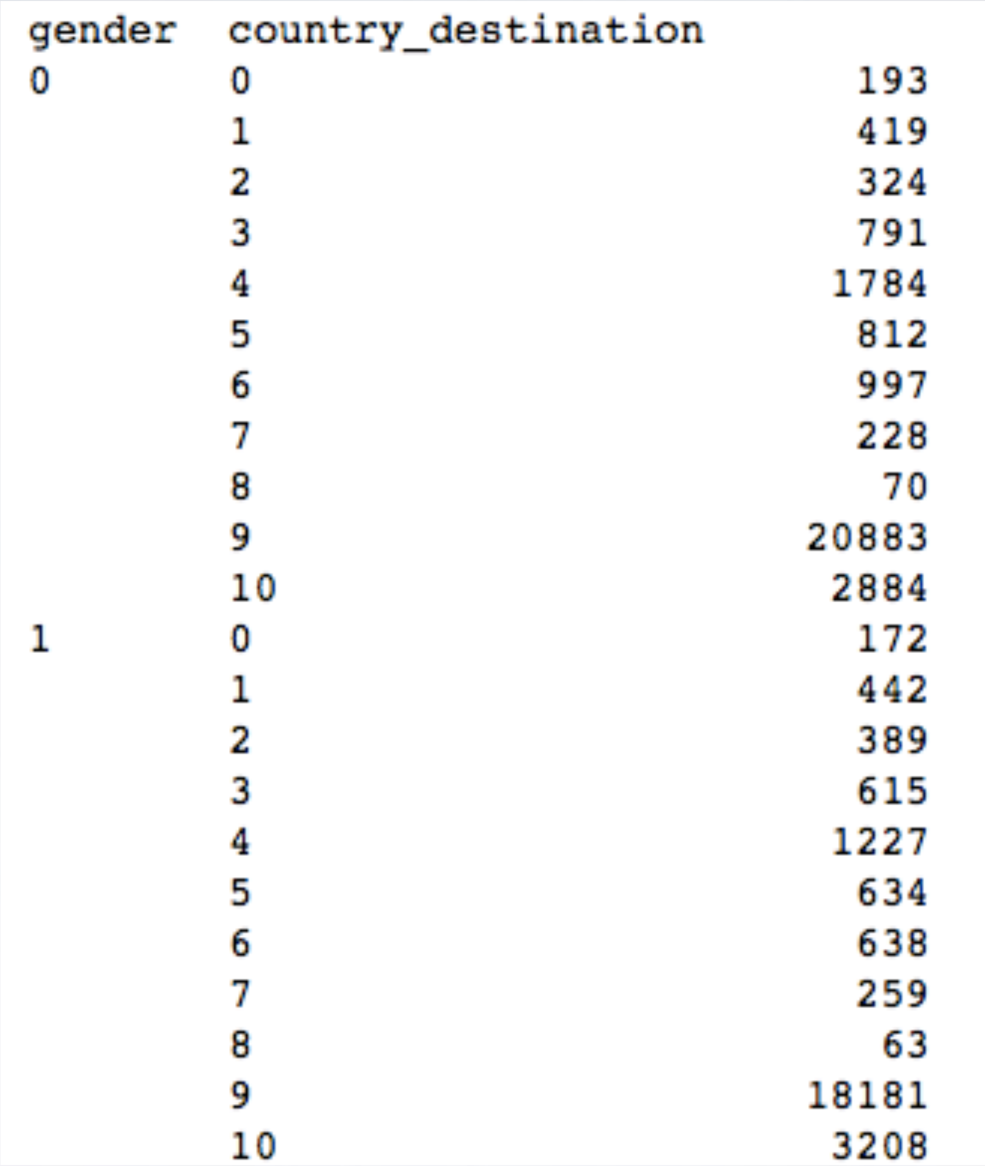


AGE DISTRIBUTION

x-axis is age. Ages below 14 and older than 100 were dropped. Blue is women, yellow is men.

COUNTRIES

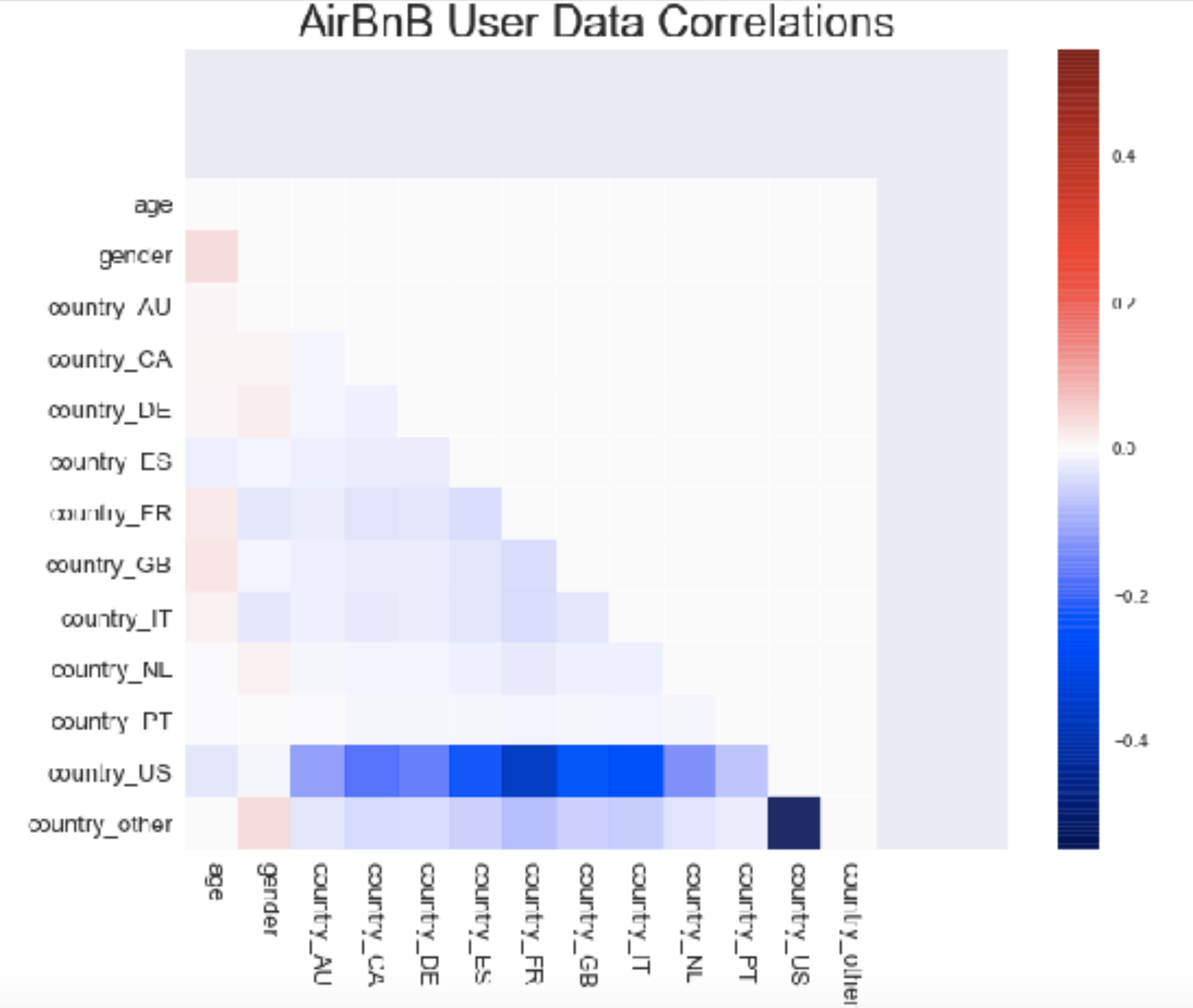
- Assume users are in the US
- Country set is 9 countries, and “other”
- Most people visit the United States



```
gender  country_destination
0       0                         193
        1                         419
        2                         324
        3                         791
        4                        1784
        5                         812
        6                         997
        7                         228
        8                          70
        9                       20883
        10                       2884
1       0                         172
        1                         442
        2                         389
        3                         615
        4                        1227
        5                         634
        6                         638
        7                         259
        8                          63
        9                       18181
        10                       3208
```

GENDER & COUNTRY

As you can see, the difference by gender is not a lot. Most people go to the US (#9). Second most France (#4).

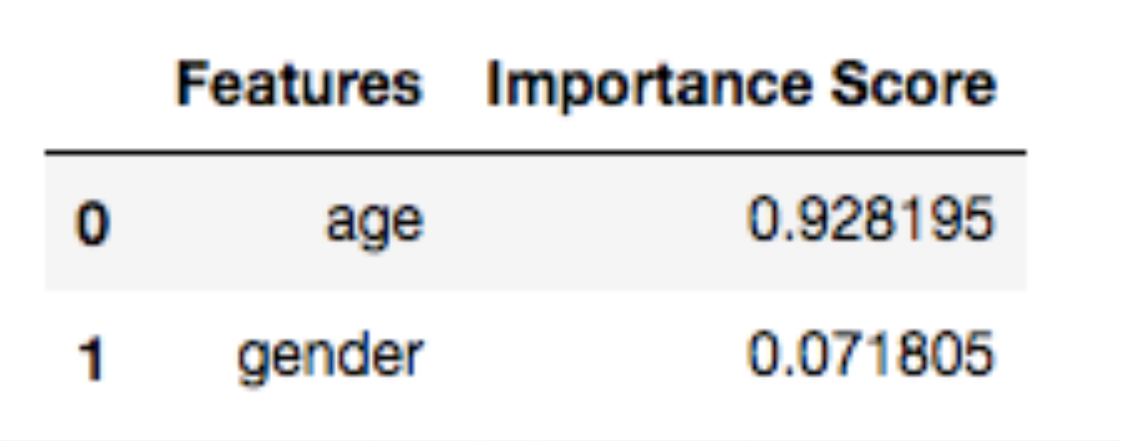


CORRELATIONS

This showed age and gender weren't strongly correlated two my destination variables. However the US appears to be very correlated with other destinations.

MODEL

RANDOM FOREST



	Features	Importance Score
0	age	0.928195
1	gender	0.071805

AGE WAS A MUCH BETTER PREDICTOR

This corresponds with what we saw in the data exploration.

MEAN SQUARE ERROR

5.35 - So there is room for improvement…

WHY?

CARTOONSTOCK
.com
Search ID: bfrn364

"I'M JUST SAYING, IF OUR INCOME IS 'FIXED', HOW COME WE'RE 'BROKE'?"

NEXT STEPS

- Try with random forest with GridSearch
- Try with LightGBM
- Look at other variables that could be good predictors

THANK YOU

APPENDIX

WHAT’S IN THE DATA?

- test_users.csv - the test set of users
    - id: user id
    - date_account_created: the date of account creation
    - timestamp_first_active: timestamp of the first activity, note that it can be earlier than date_account_created or date_first_booking because a user can search before signing up
    - date_first_booking: date of first booking
    - gender
    - age
    - signup_method
    - signup_flow: the page a user came to signup up from
    - language: international language preference
    - affiliate_channel: what kind of paid marketing
    - affiliate_provider: where the marketing is e.g. google, craigslist, other
    - first_affiliate_tracked: whats the first marketing the user interacted with before the signing up
    - signup_app
    - first_device_type
    - first_browser
    - country_destination: this is the target variable you are to predict

WHAT'S IN THE DATA?

- sessions.csv - web sessions log for users
    - user_id: to be joined with the column 'id' in users table
    - action
    - action_type
    - action_detail
    - device_type
    - secs_elapsed
- countries.csv - summary statistics of destination countries in this dataset and their locations
- age_gender_bkts.csv - summary statistics of users' age group, gender, country of destination