AIRBNB NEW USER BOOKING CAPSTONE PROJECT

PROJECT PROCESS

- Project Objective Definition
- Data Cleaning
- Data Exploratory Analysis
- In-Depth Analysis: Classification modeling
- Modeling Performance Validation
- Presentation

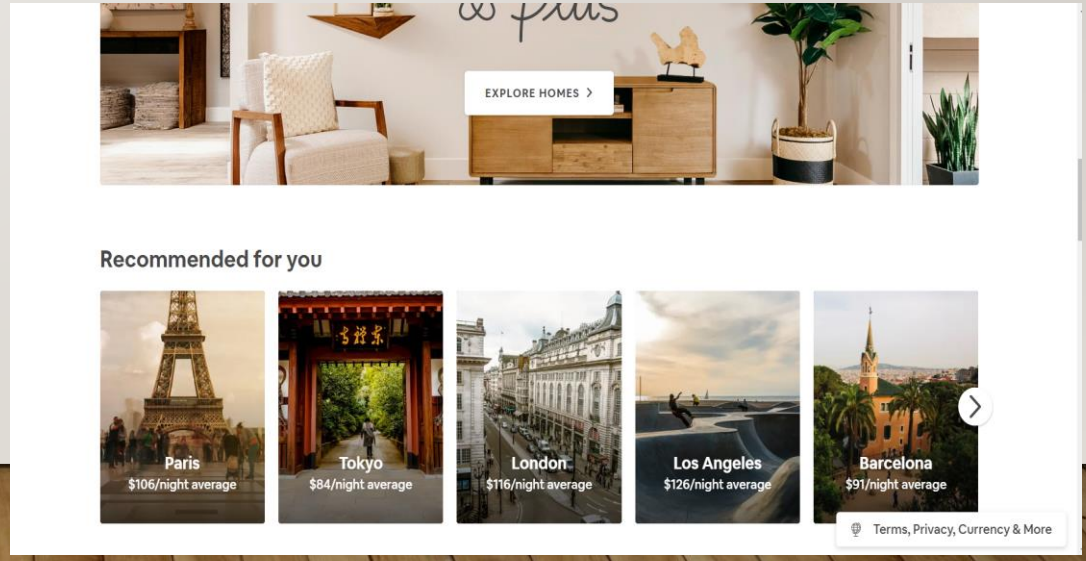
PROJECT BACKGROUND & OBJECTIVE

- Kaggle Competition: Where will a new guest book their first travel experience?
- Predict the first destination country for the new users
- Deliver customized content to user to increase user booking rate

Airbnb New User Bookings

Where will a new guest book their first travel experience?

1,462 teams · 3 years ago



DATA

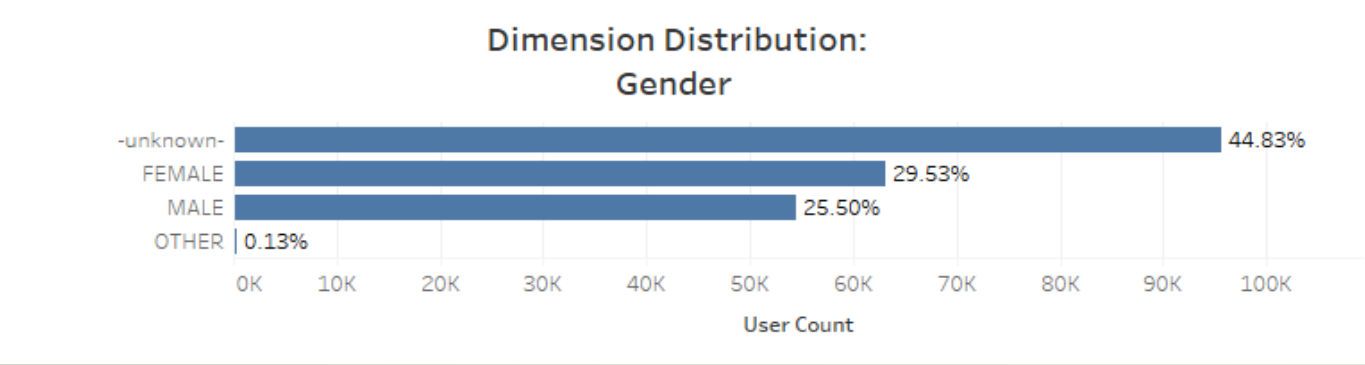
- User demographics
  - Gender
  - Age
  - Sign-up method
  - Language
- Web session records
- Destination country: US, FR, CA, GB, ES, IT, PT, NL,DE,AU, NDF(Not booked yet) and Other

AIRBNB NEW USER BOOKING CAPSTONE PROJECT

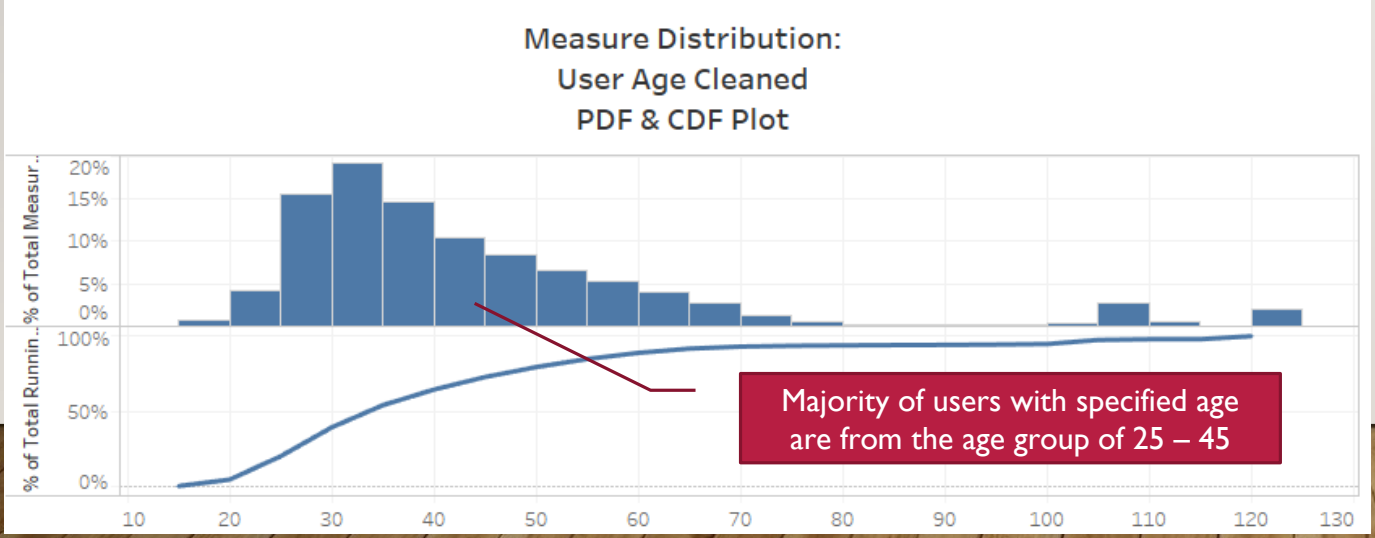
EXPLORATORY ANALYSIS

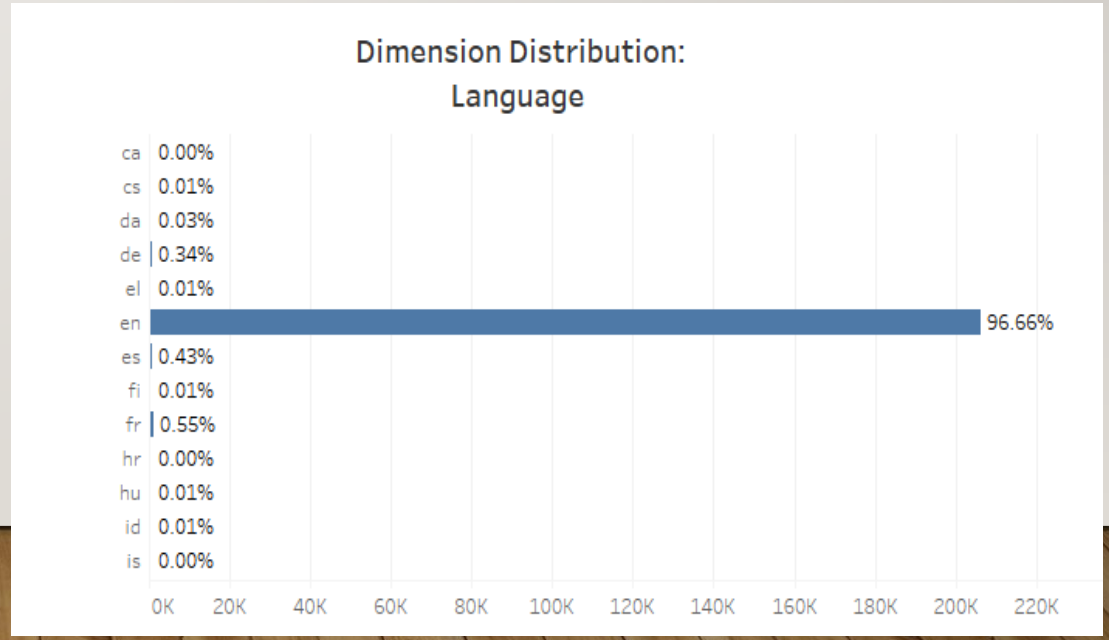
USER DEMOGRAPHIC

WHAT DO THE USERS IN THE DATASET LOOK LIKE ?



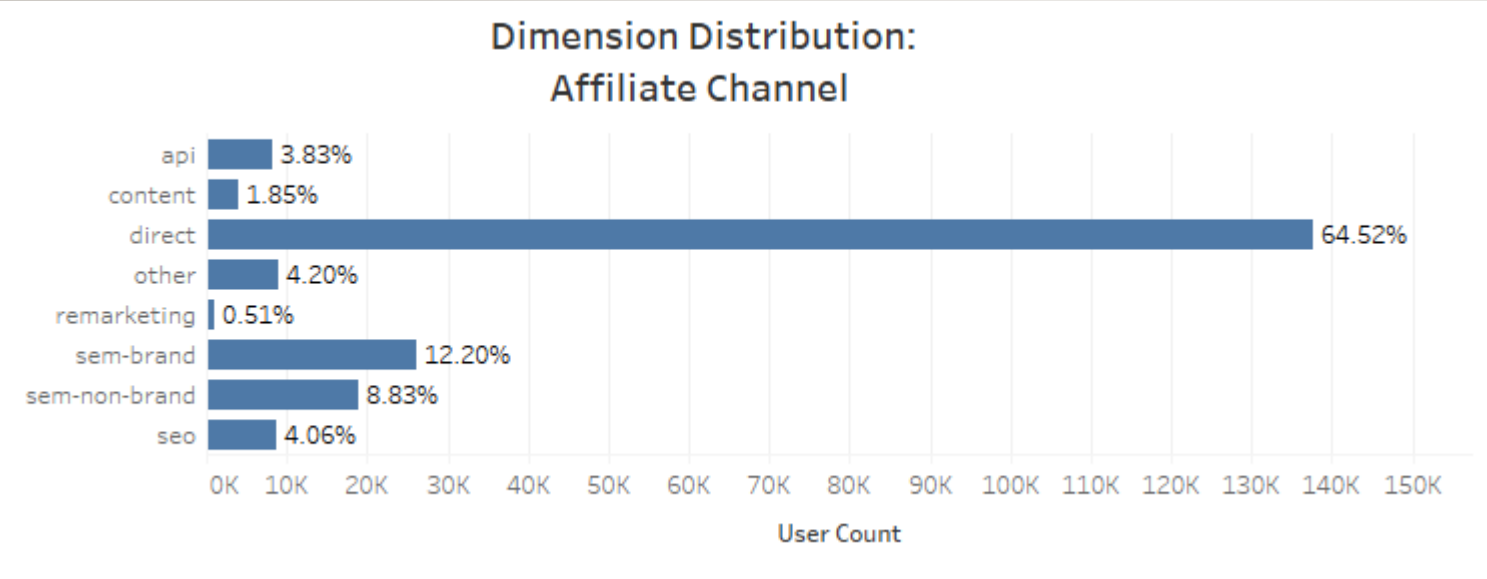
- 213,451 users in total from US.
- 45% of users have unknown gender.
- 59% of users have unknown age.





USER ACQUISITION CHANNELS

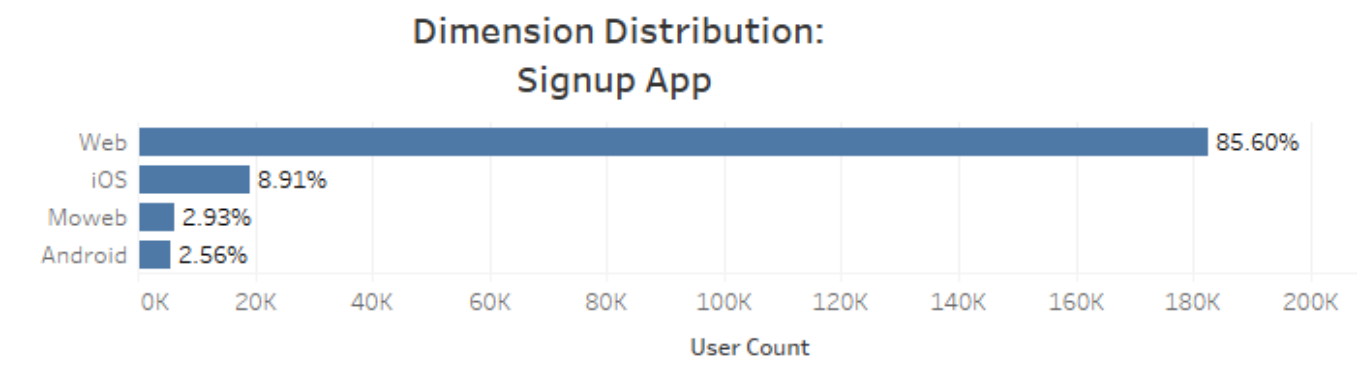
WHERE DO THE USERS COME FROM ?

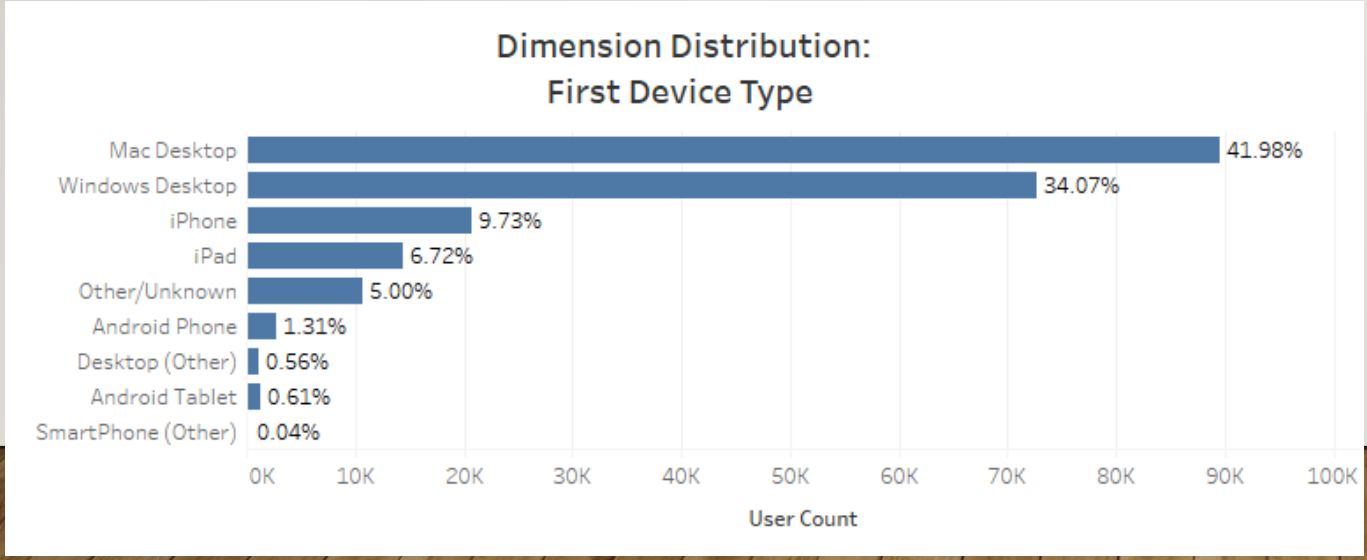


- Most users are either coming to the website by typing in URL directly or coming from Google.

USER BEHAVIORS

HOW DID THE USERS COME TO THE WEBSITE FOR THE FIRST TIME ?

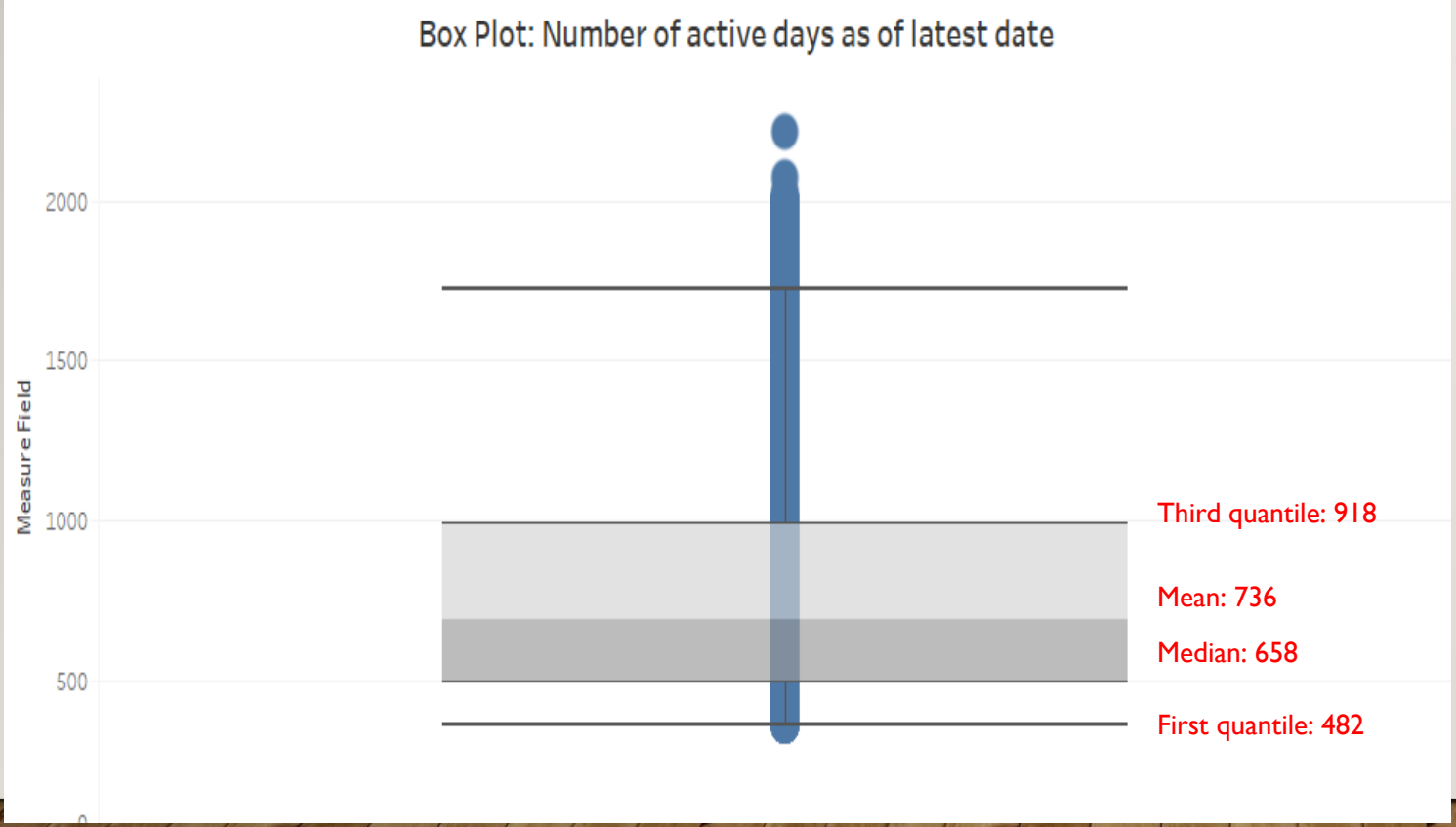




- Majority of users signed up through desktop
- Mac is more popular among users than any other device

USER BEHAVIORS

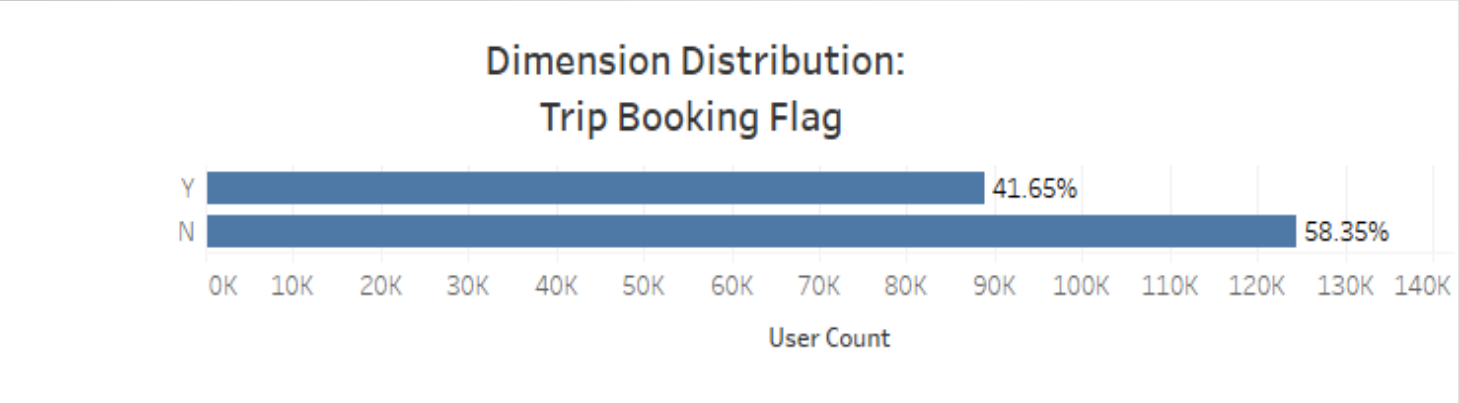
HOW LONG HAVE USERS BEEN BROWSING ON THE WEBSITE?

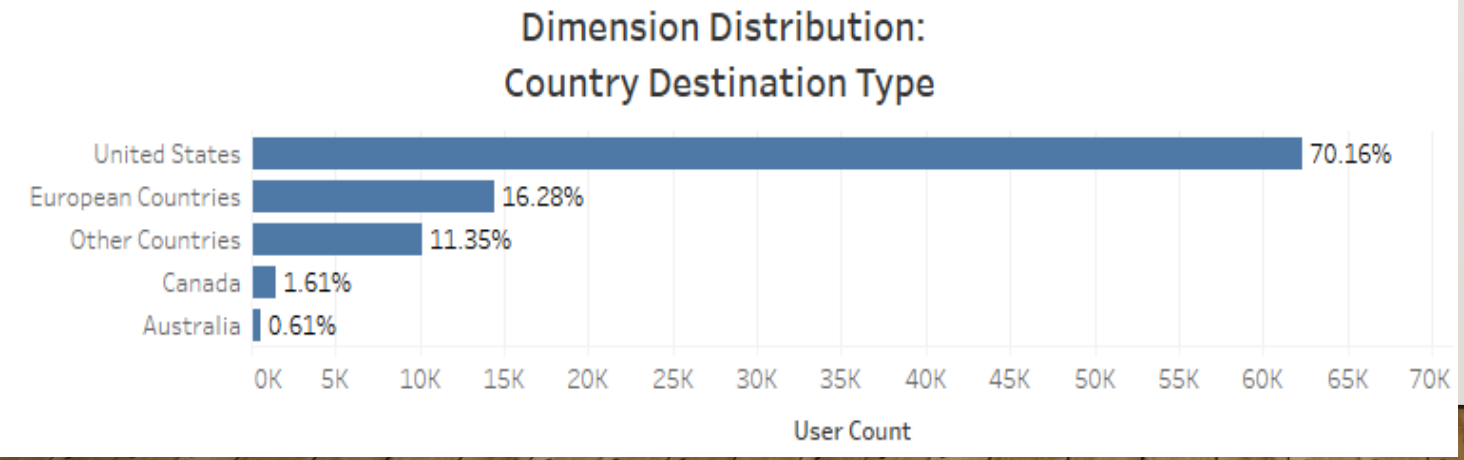


- As of 7/1/2015, user age ranges from 1 year to a bit over 6 years. Most of users are 1 to 2.5 years old based on their first active date on the site.

BOOKING HISTORY

HOW MANY USERS HAVE MADE THE BOOKING ? WHAT DESTINATION ARE MORE POPULAR ?

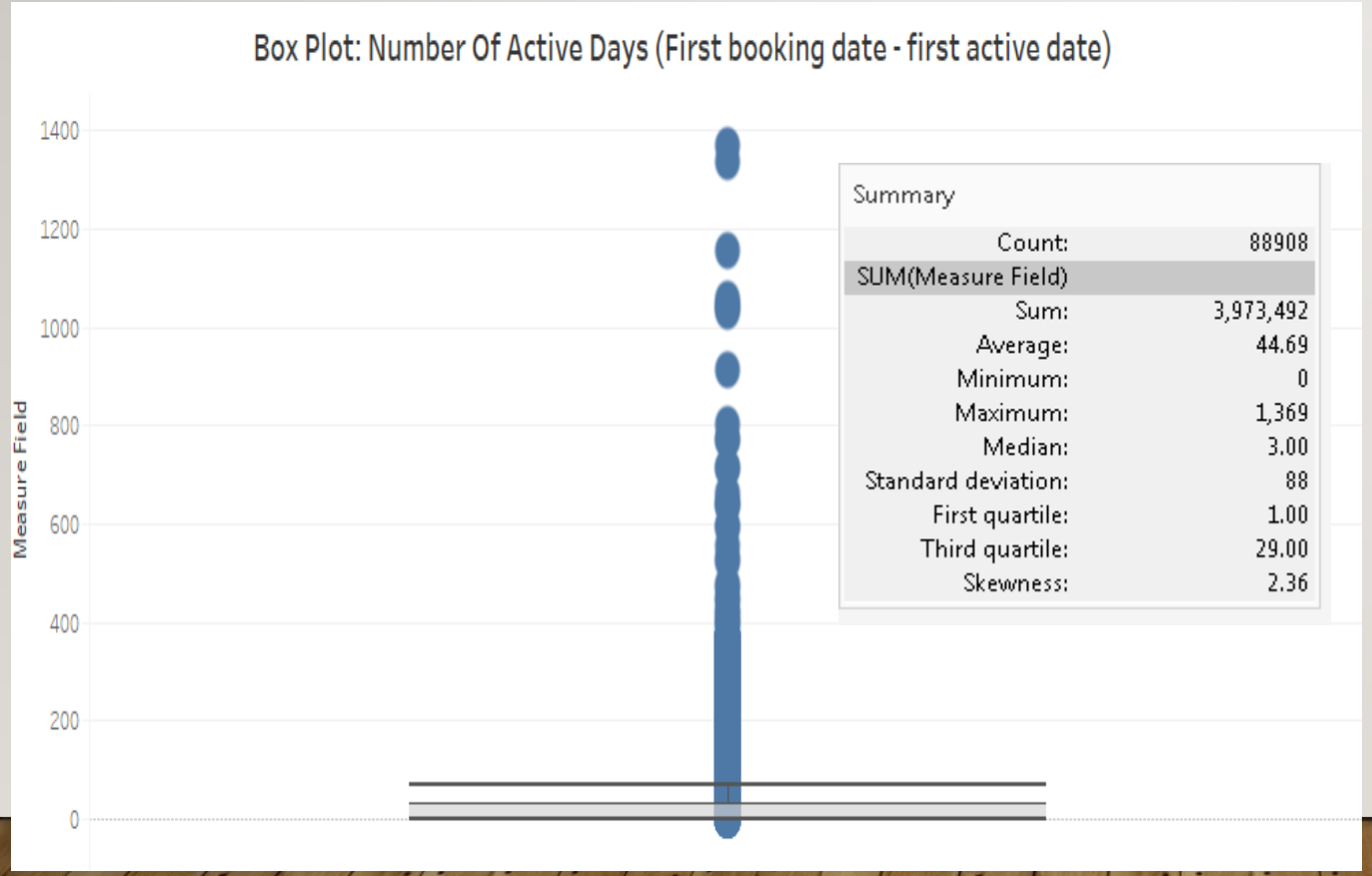




- Only 42% of users have made a booking.
- Most users like to travel within US. Europe is the second place they like to go.

BOOKING HISTORY

HOW LONG DID IT TAKE USERS TO COVERT AFTER GETTING ON THE SITE?

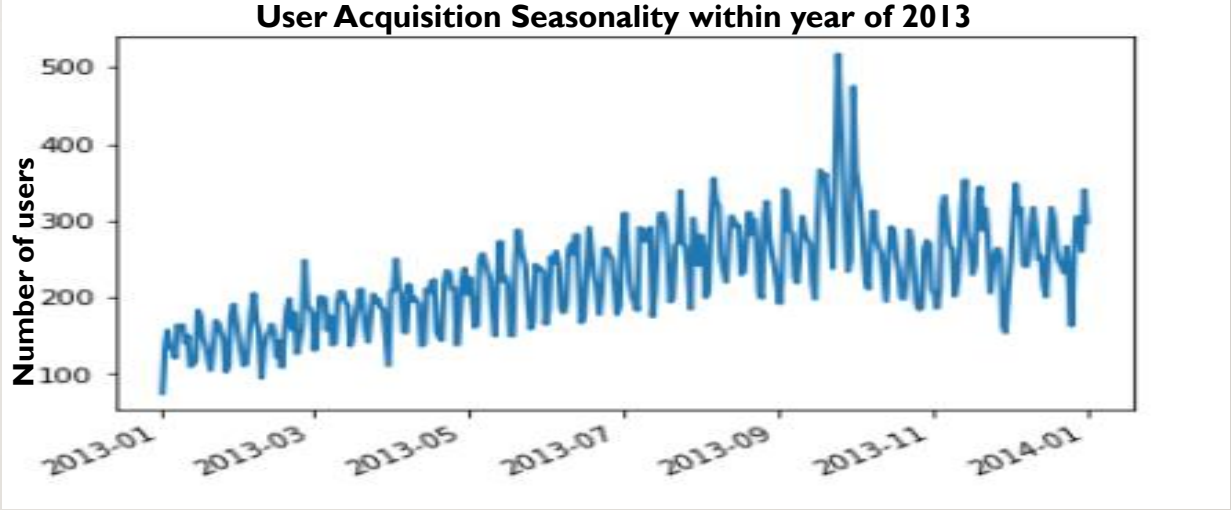


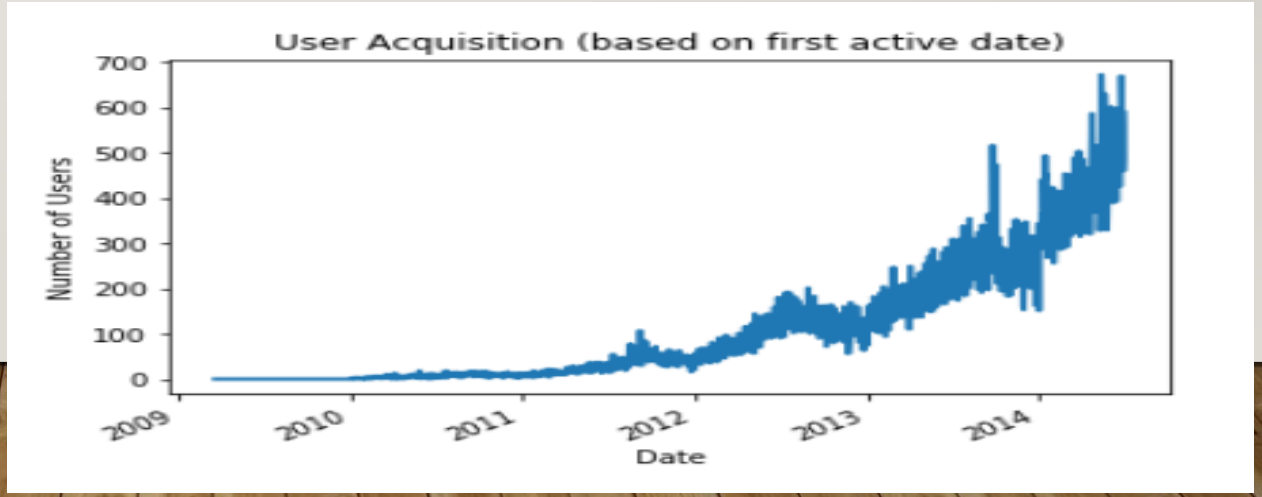
- Majority of bookers (75%) have booked their first destination within 30 days since they got on the website for the first time.

BOOKING HISTORY

HOW DOES THE USER GROWTH FOR THE WEBSITE LOOK?

- Between 2013 and 2014, number of users on Airbnb has grown dramatically.
- There's seasonality for the user acquisition:
  - October is peak month for user acquisition.
  - The beginning of the year is the off-peak season.

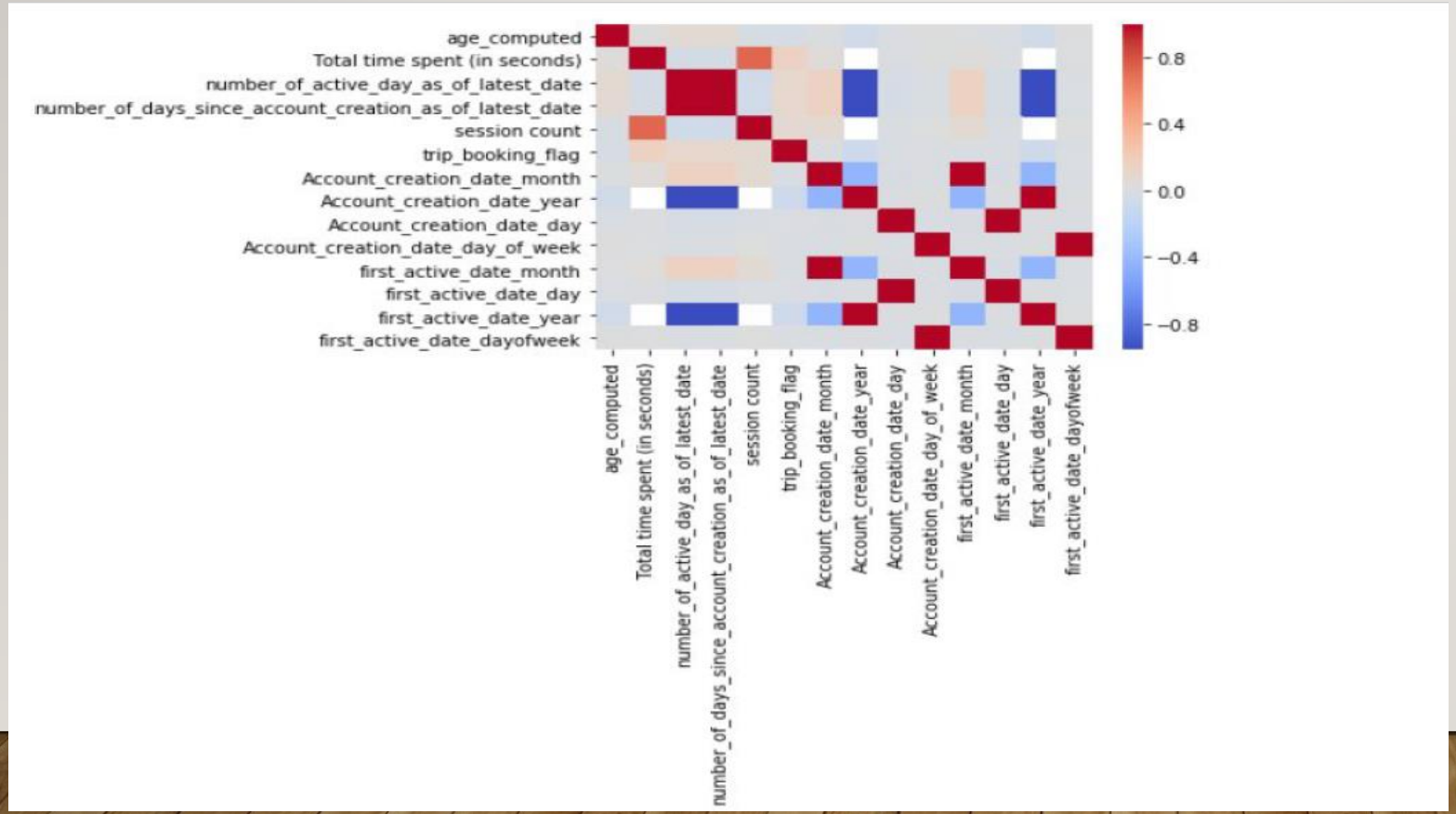




CORRELATION EXPLORATION

IS THERE ANY CORRELATION BETWEEN USER PROFILE & BEHAVIORS AND TRIP BOOKING ?

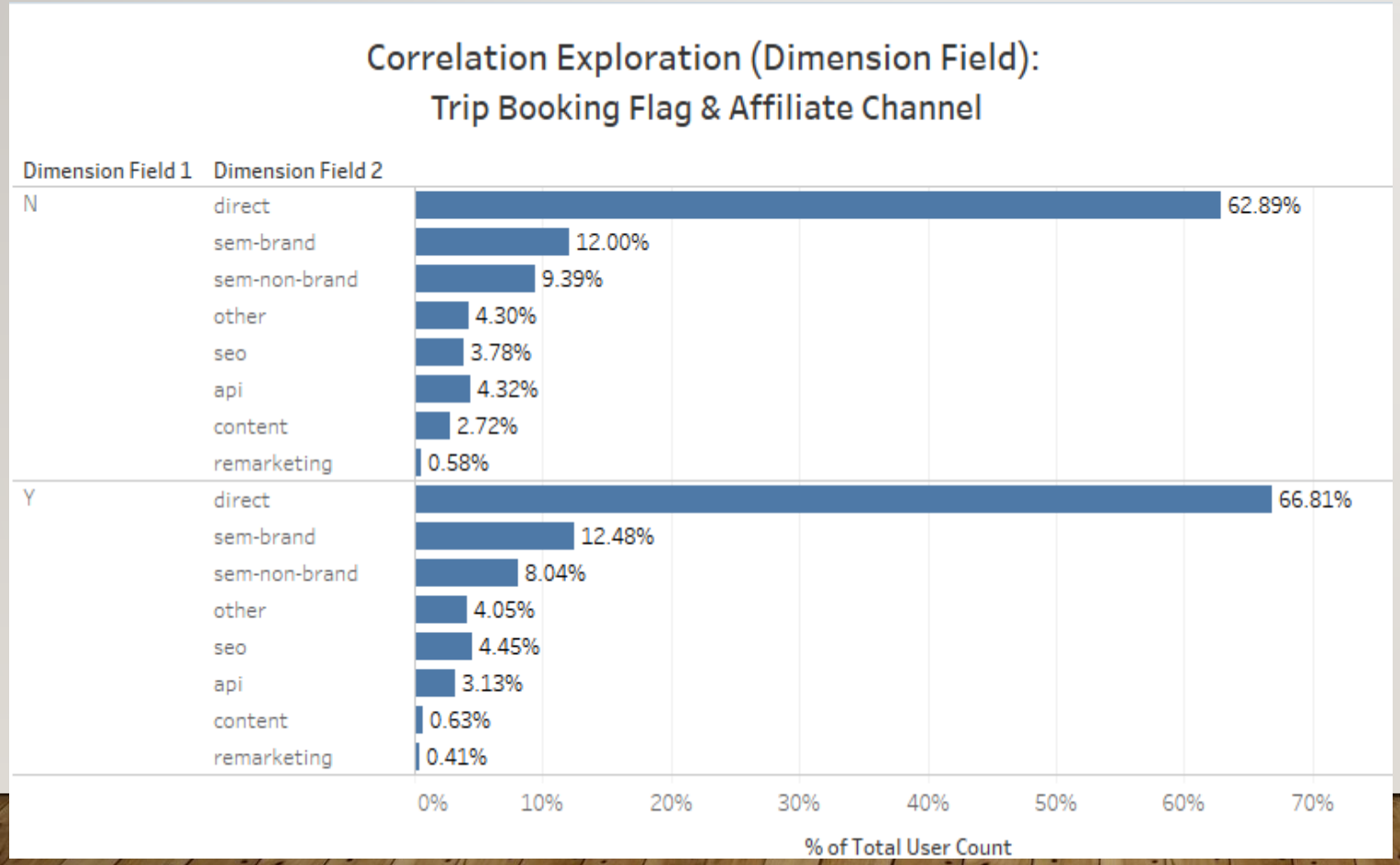
CORRELATION HEATMAP



- Total time spent on the Airbnb site is more correlated to the user booking.

CORRELATION EXPLORATION

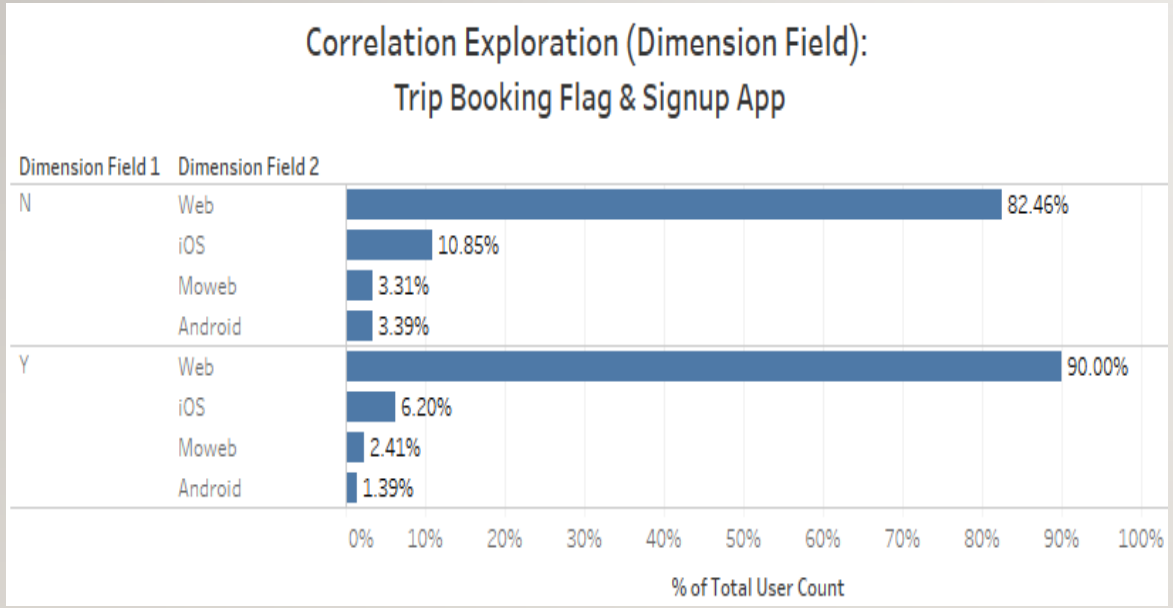
IS THERE ANY CORRELATION BETWEEN USER ACQUISITION CHANNEL AND THEIR BOOKING BEHAVIOR ?

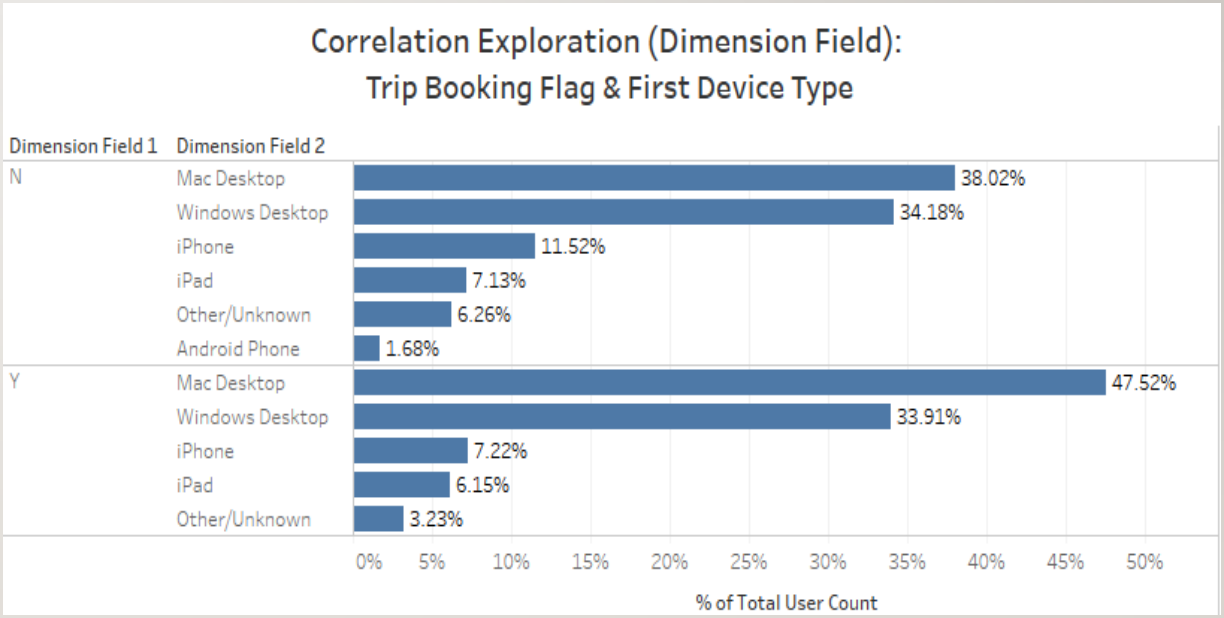


- Users coming directly to the website are slightly more like to make a booking.

CORRELATION EXPLORATION

IS THERE ANY CORRELATION BETWEEN USER ACQUISITION CHANNEL AND THEIR BOOKING BEHAVIOR ?

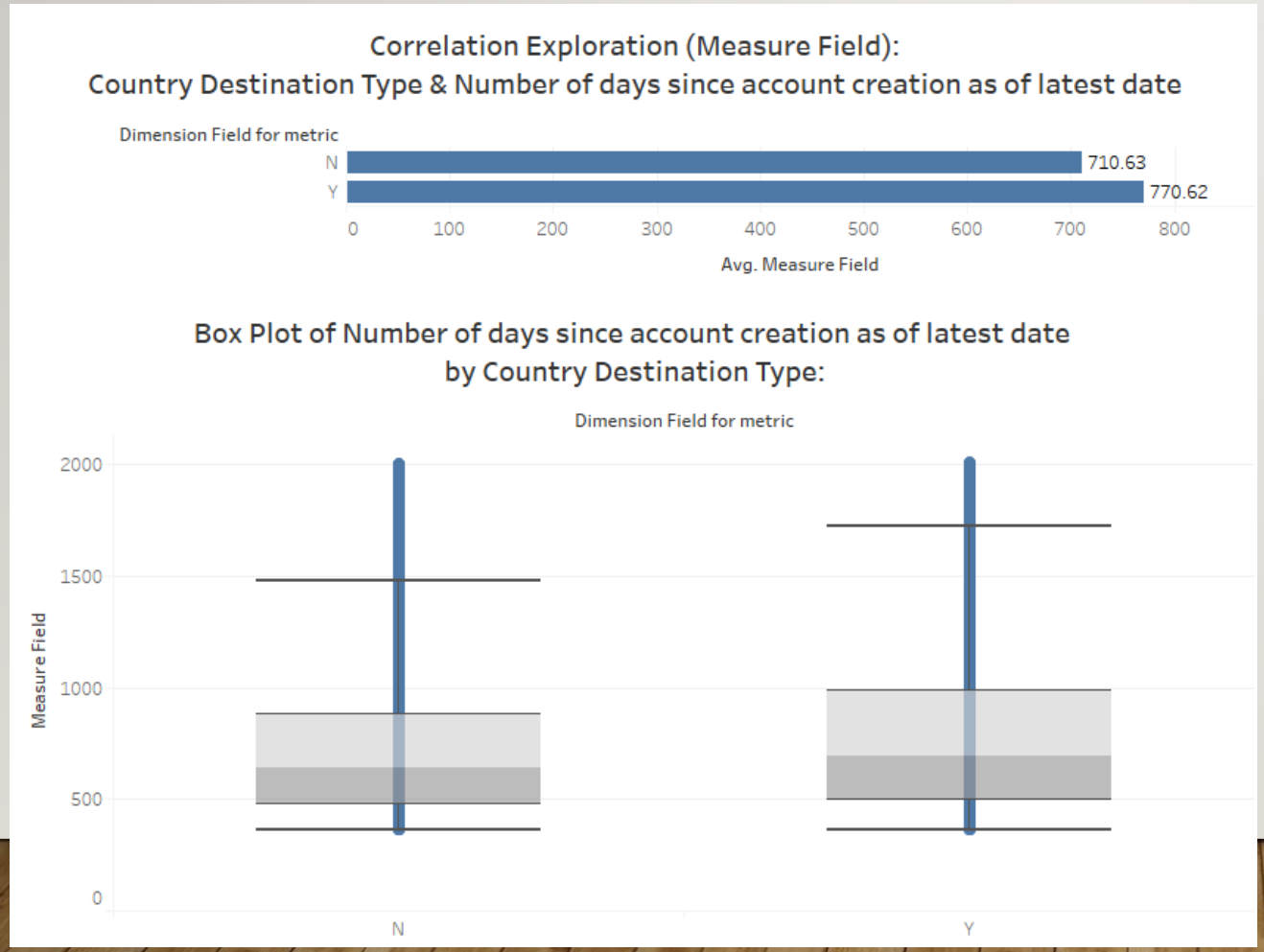




- Users coming from Web are more likely to make a booking.
- Users signing up using Mac are more likely to make a booking.

CORRELATION EXPLORATION

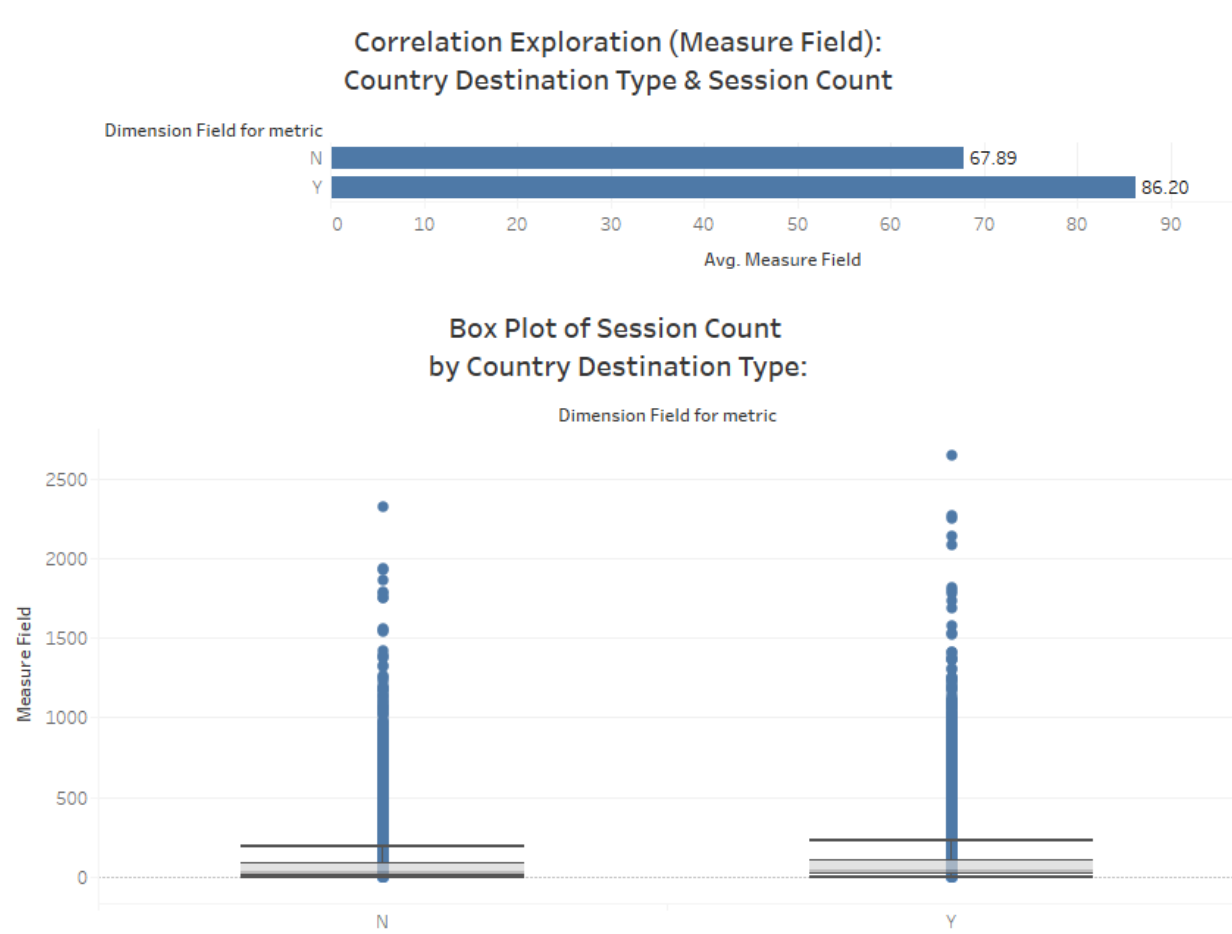
IS IT TRUE THE MORE TIME USERS HAVE SPENT WITH WEBSITE THE MORE LIKELY USERS ARE TO MAKE A BOOKING ?



- Users that have made a booking have longer account age on average than users that have not made a booking.

CORRELATION EXPLORATION

IS IT TRUE THE MORE TIME USERS HAVE SPENT WITH WEBSITE THE MORE LIKELY USERS ARE TO MAKE A BOOKING ?



- Users that have made a booking have done more activities and spent more time with the website on average than users that have not made a booking.

HYPOTHESIS TESTING ON USER BOOKING BEHAVIORS

- Hypothesis 1: Bookers has stayed with the website longer than the non-bookers (based on the first active date).
  - Conclusion: Significant difference between bookers and non-bookers in number of active days.
- Hypothesis 2: Bookers have more activities on the website than non-bookers (based on the session count).
  - Conclusion: Bookers are on average more active than non-bookers on the website.
- Hypothesis 3: Age and gender could be important factors to predict which country users booked.
  - Conclusion: Significant relationship between age, gender and booking destination. But the relationship is not that strong.

AIRBNB NEW USER BOOKING CAPSTONE PROJECT

IN-DEPTH ANALYSIS

TWO STEP CLASSIFICATION

FIRST STEP: BINARY CLASSIFICATION

- Label: Trip booking flag (if a user has booked or not)
- Features: One-hot encoding to create dummy variables
- Train test split (70% training, 30% testing data)
- Metric: Accuracy score, precision, recall, Roc Auc Score

SCREENSHOT OF SOME FEATURES

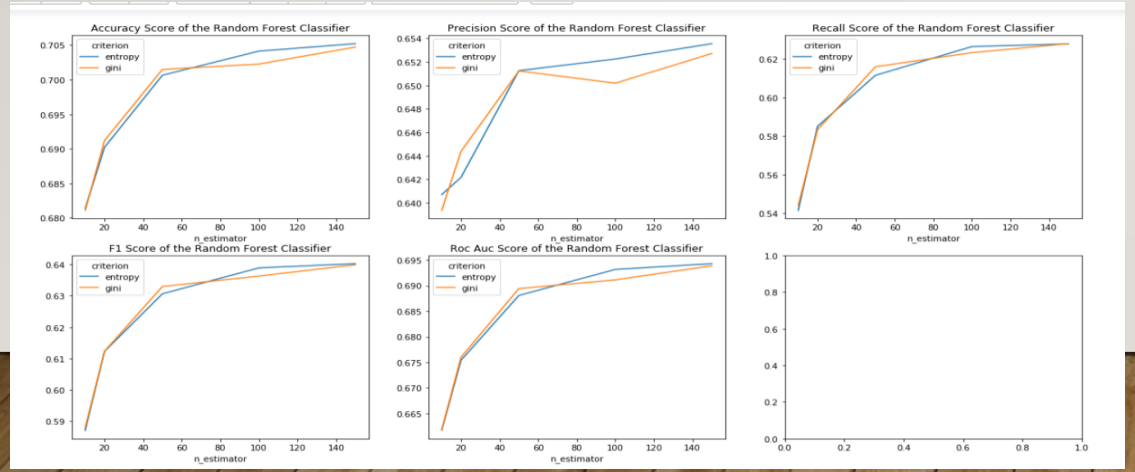
```
Index(['id', 'date_account_created', 'timestamp_first_active_cleaned',
       'gender', 'signup_method', 'signup_flow', 'language',
       'affiliate_channel', 'affiliate_provider', 'first_affiliate_tracked',
       'signup_app', 'first_device_type', 'first_browser',
       'country_destination', 'age_computed',
       'Account_creation_before_booking_flag', 'Total time spent (in seconds)',
       'number_of_active_day_as_of_latest_date',
       'number_of_days_since_account_creation_as_of_latest_date',
       'session count', 'trip_booking_flag', 'Account_creation_date_month',
       'Account_creation_date_year', 'Account_creation_date_day',
       'Account_creation_date_day_of_week', 'first_active_date_month',
       'first_active_date_day', 'first_active_date_year',
       'first_active_date_dayofweek', 'age_bucket',
       'Total time spent (in seconds)_fill_null_zero',
       'session count_fill_null_zero', 'Total time spent (in seconds)_bucket',
       'session count_bucket'],
      dtype='object')
```

TWO STEP CLASSIFICATION

FIRST STEP: BINARY CLASSIFICATION

- Models used: Logistic regression, random forest , XG Boost
- Hyperparameter Tuning on parameters like number of estimators via Randomized Search and Grid Search
- Optimization: Random forest has the best performance on both training and testing set compared with other models.

HYPERPARAMETER TUNING FOR RANDOM FOREST CLASSIFIER



PERFORMANCE OF RANDOM FOREST CLASSIFIER

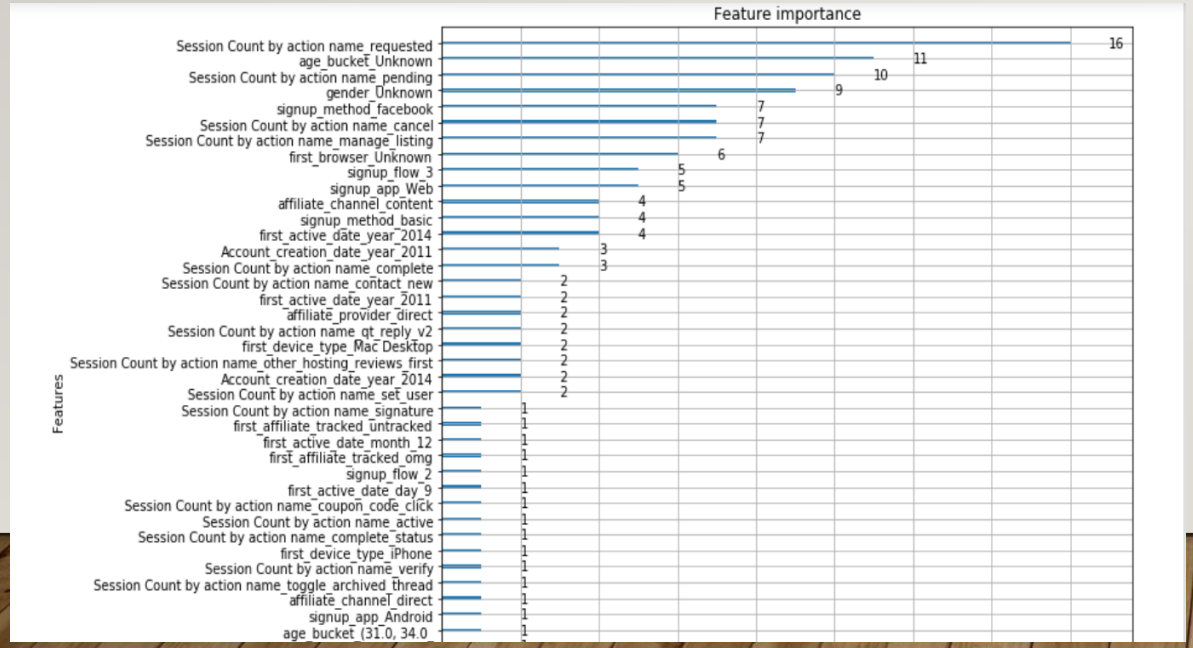
```
In [50]:  1  model_evaluation(best_grid, X_train, y_train)

Model Performance:
accuracy score:  0.7845806980557508
precision score:  0.7665780017732978
recall score:  0.6935125246270757
roc auc score: 0.7715004950116321

In [51]:  1  model_evaluation(best_grid, X_test, y_test)

Model Performance:
accuracy score:  0.7243447096577733
precision score:  0.6865488463426608
recall score:  0.6268489466606902
roc auc score: 0.7106188540412093
```

TWO STEP CLASSIFICATION

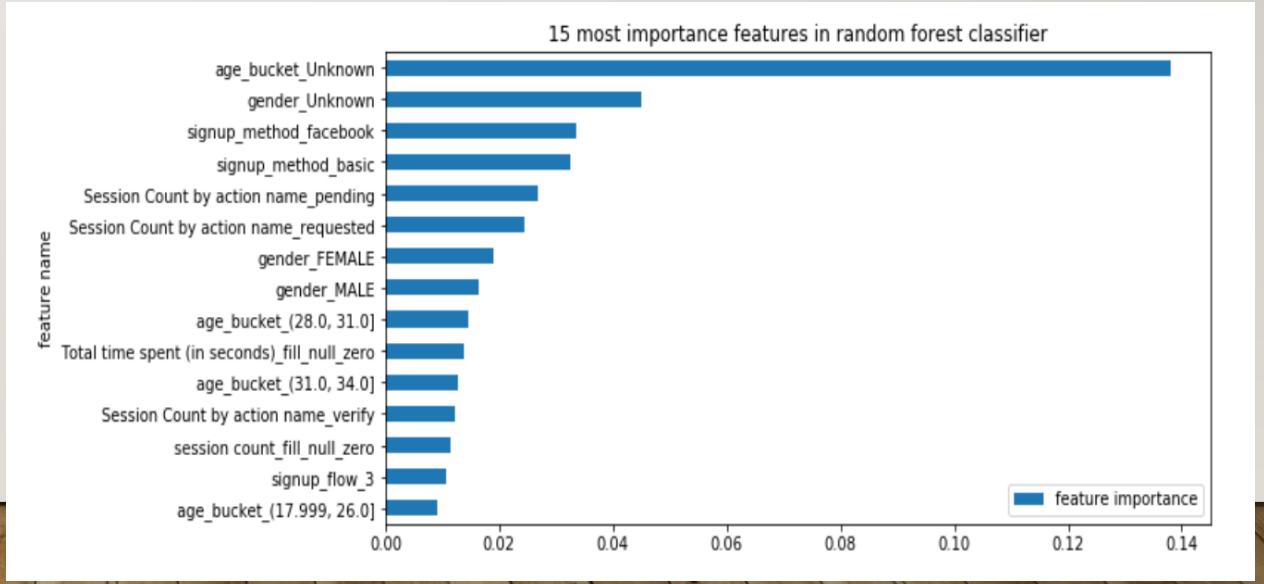
FIRST STEP: BINARY CLASSIFICATION

- Model interpretation from feature importance:
  - Random Forest model has given more weight on demographic information of the users.
  - Xgboost has put more weight on types of activities users have done on the website.

FEATURE IMPORTANCE OF XGBOOST



FEATURE IMPORTANCE OF RANDOM FOREST

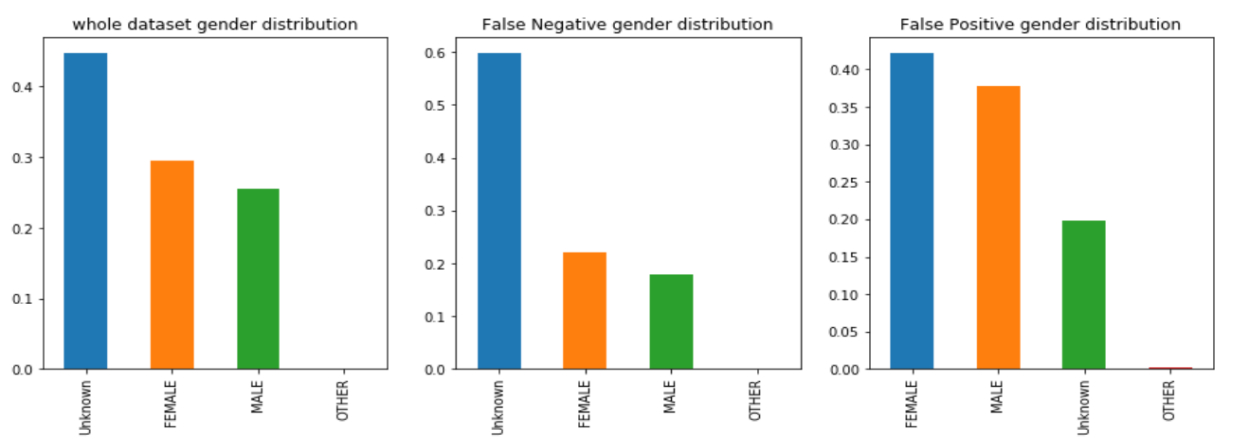


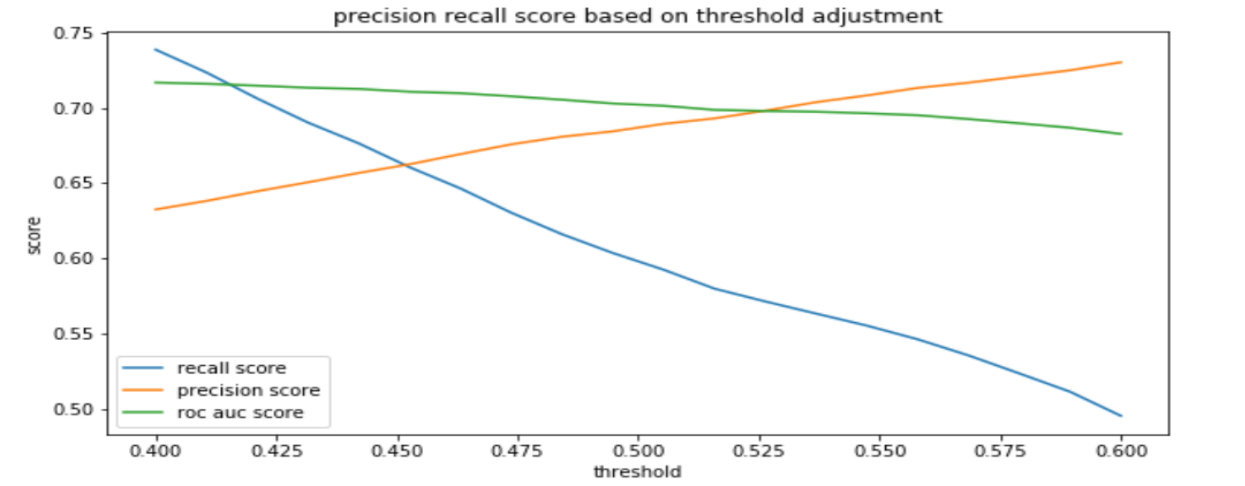
TWO STEP CLASSIFICATION

FIRST STEP: BINARY CLASSIFICATION

- Error analysis
  - Users with unknown gender or age are less likely to classified as bookers.
  - Users using basic signup method are more likely to be classified as bookers.
  - Users that signed up through Facebook are more likely to be classified as non-bookers.
- Threshold Adjustment
  - Use 0.44 instead 0.55 as the probability threshold since it gives a good balance between precision and recall.

ERROR ANALYSIS





TWO STEP CLASSIFICATION

SECOND STEP: MULTI CLASS CLASSIFICATION

- Label: Destination Country
- Features: same as features for binary classification
- Metric: Accuracy score and NDCG score

NDCG SCORE

The evaluation metric for this competition is NDCG (Normalized discounted cumulative gain) @k where k=5. NDCG is calculated as:

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2 (i + 1)},$$

$$nDCG_k = \frac{DCG_k}{IDCG_k},$$

where rel_i is the relevance of the result at position i.

$IDCG_k$ is the maximum possible (ideal) DCG for a given set of queries. All NDCG calculations are relative values on the interval 0.0 to 1.0.

For each new user, you are to make a maximum of 5 predictions on the country of the first booking. The ground truth country is marked with relevance = 1, while the rest have relevance = 0.

For example, if for a particular user the destination is FR, then the predictions become:

[FR] gives a $NDCG = \frac{2^1-1}{log_2(1+1)} = 1.0$

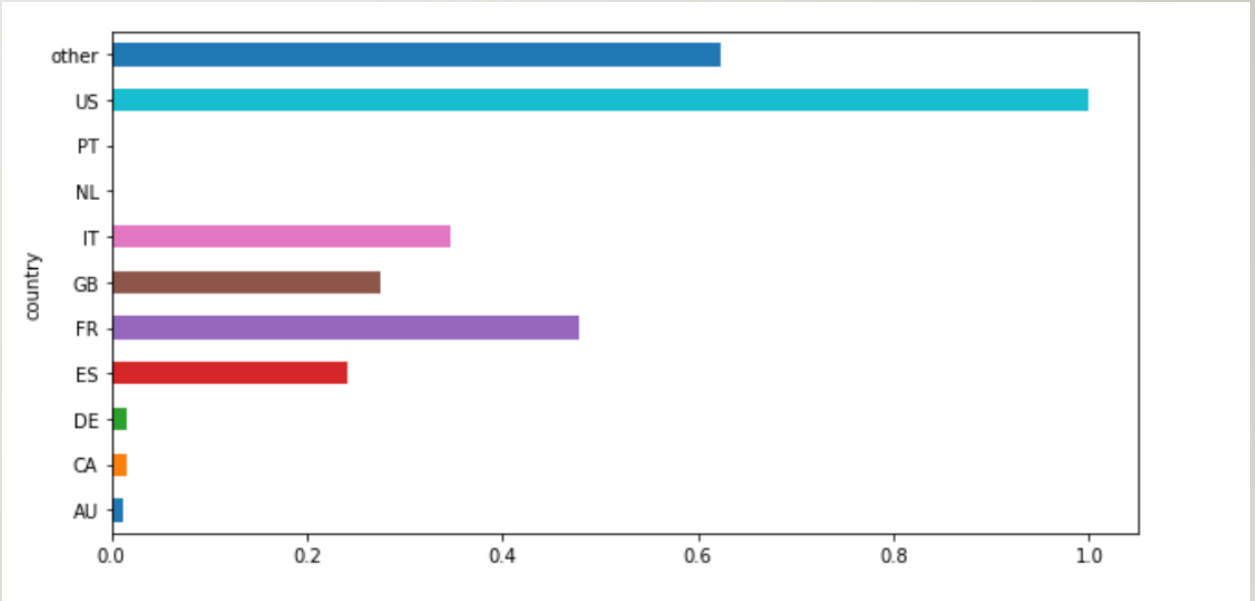
[US, FR] gives a $DCG = \frac{2^0-1}{log_2(1+1)} + \frac{2^1-1}{log_2(2+1)} = \frac{1}{1.58496} = 0.6309$

TWO STEP CLASSIFICATION

SECOND STEP: MULTI CLASS CLASSIFICATION

- Models used: Random forest, logistic regression and XG Boost.
- Optimization: XG Boost has the best result with accuracy score of 70% and average NDCG score of 82% on the testing set.

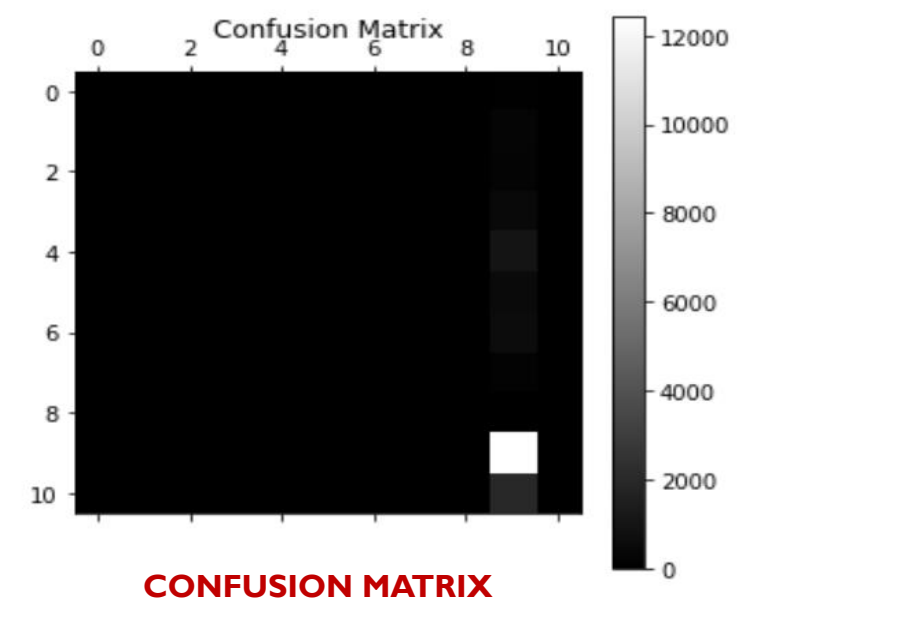
AVERAGE NDCG SCORE BY COUNTRY



TWO STEP CLASSIFICATION

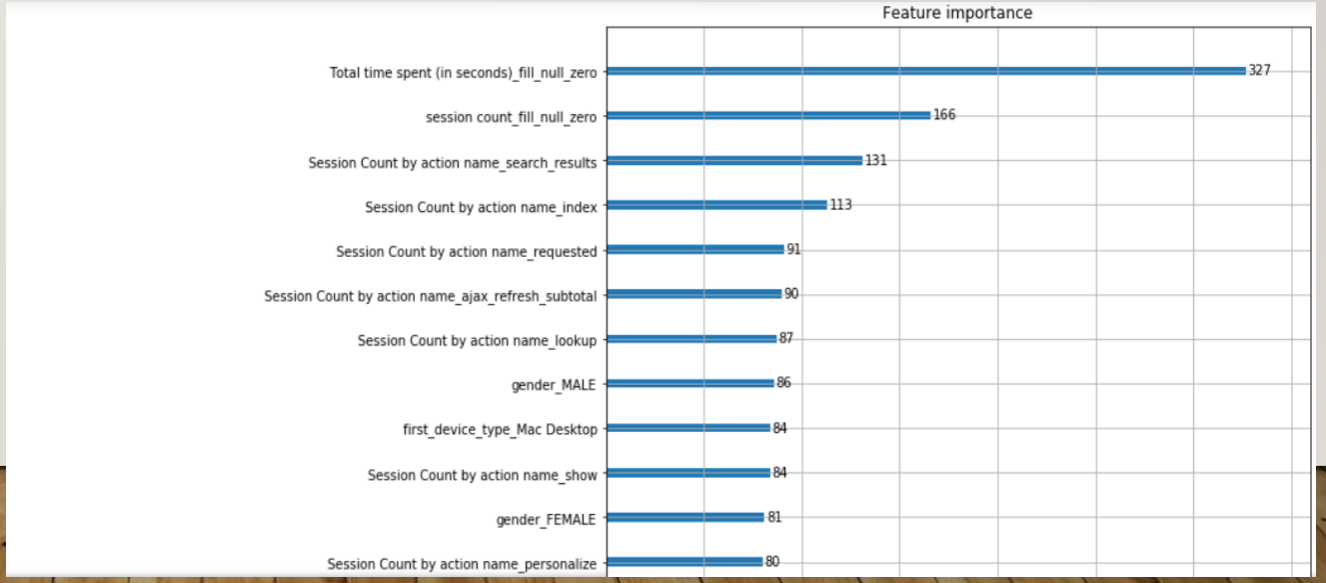
SECOND STEP: MULTI CLASS CLASSIFICATION

- Error analysis: The model is doing well in predicting country like US. But it's generating errors by misclassifying some of the users who booked other countries into US.
- Feature importance: Some important features for XG boost model include total time spent, session count, search result session count, index session count and requested session count.



CONFUSION MATRIX

IMPORTANT FEATURES



NEXT STEP FOR THE PROJECT

- Fields like unknown gender and unknown age have limited the model's capability to make accurate classifications.
- There's no parameter tuning for XG Boost in the multi-class classification.
- Next step:
  - Only train on the user data with known information
  - Try hyperparameter tuning on the XG Boost model