# Airbnb New User Booking Prediction


By:

Anirudh Dave

Yilin Guan

Devang Jain

Navneet Poddar

Pallavi Varandani

# Table of Contents

# 1. Introduction

❖Given Information:

➢5 Data Frames

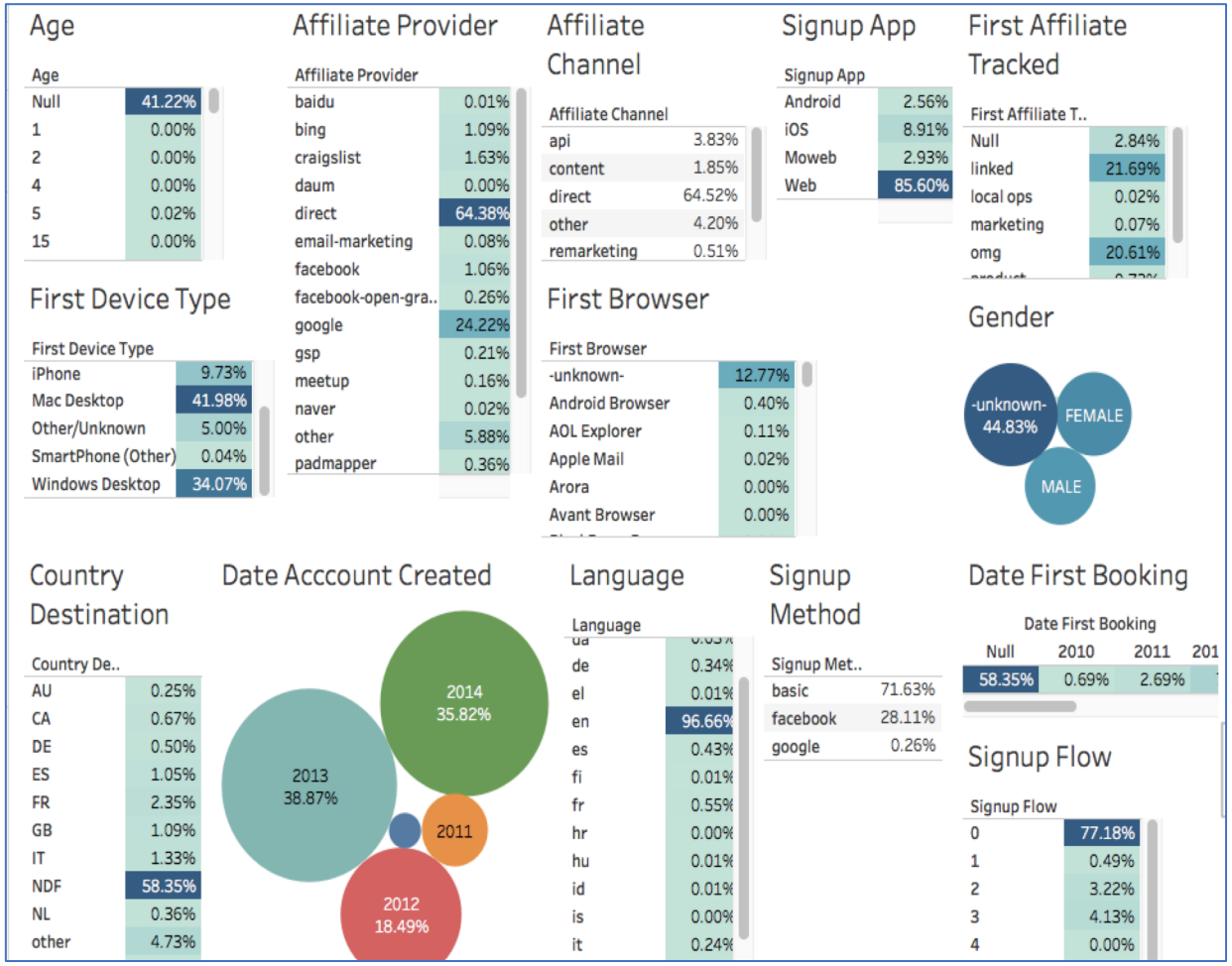
- Age Bracket Dataset:*(420 X 5)*
- Countries Dataset: *(10 X 7)*
- Training Dataset: *(213451 X 16)*
- Test Dataset: *(62096 X 16)*
- Sessions Dataset: *(10567737 X 6)*

➢MetaData

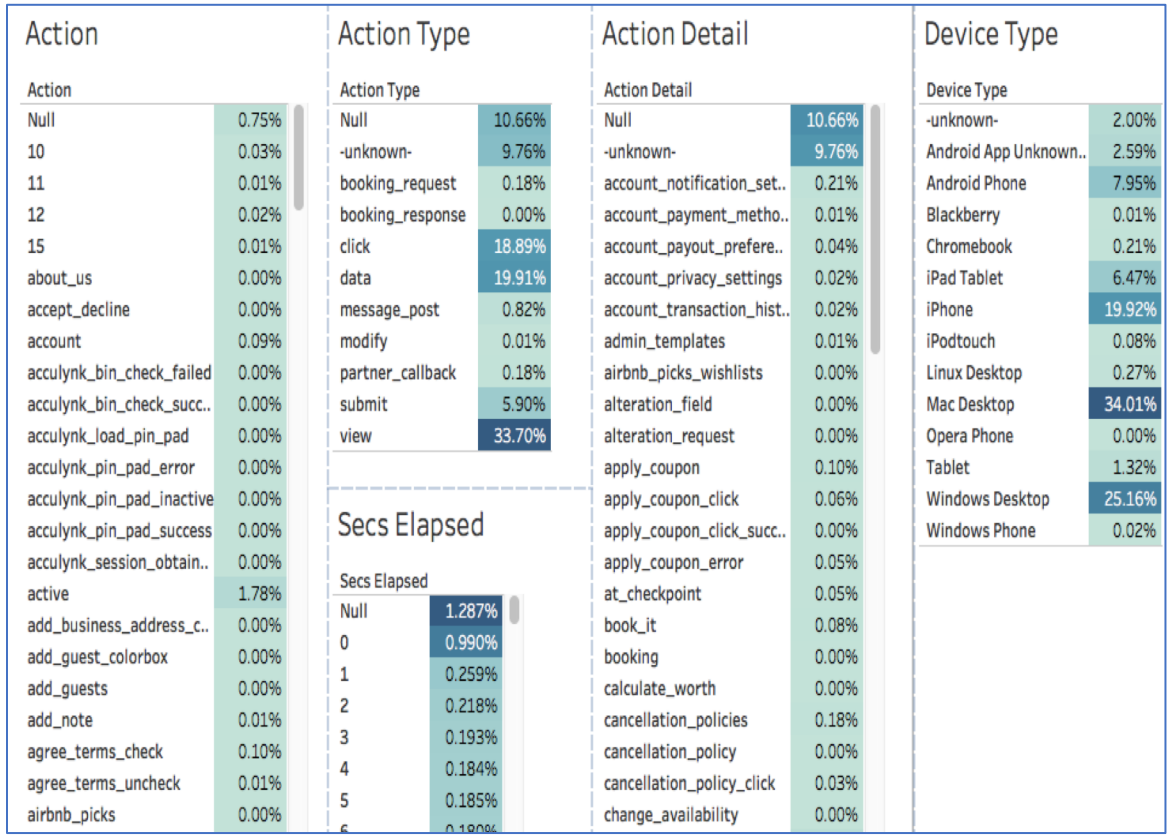
❖Goal: Predicting User's First Booking on AirBnb

# 2. Understanding Data

## TRAIN DATASET

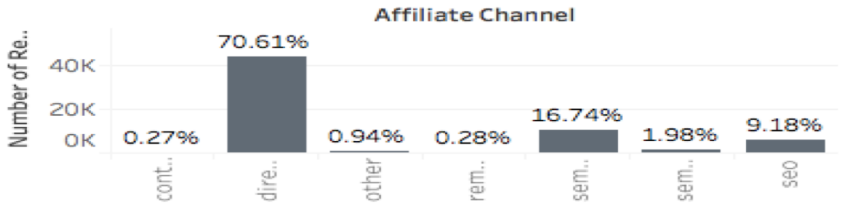



## SESSION DATASET



# TEST DATASET

## Affiliate Channel




## Affiliate Provider

| Affiliate Provid.. | |
|---|---|
| baidu | 0.00% |
| bing | 2.24% |
| craigslist | 0.01% |
| daum | 0.00% |
| direct | 70.61% |
| email-marketing | 0.17% |
| facebook | 2.77% |
| facebook-open-.. | 0.03% |
| google | 22.97% |
| gsp | 0.00% |
| meetup | 0.02% |
| naver | 0.02% |
| other | 0.78% |

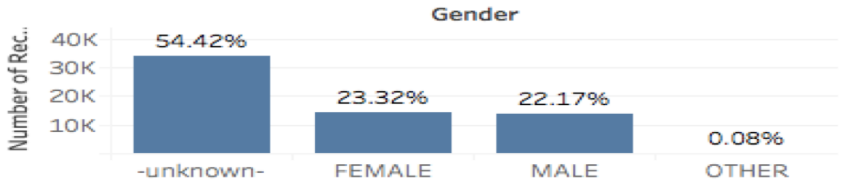
## First Browser

| First Browser | |
|---|---|
| -unknown- | 27.58% |
| Android Browser | 1.17% |
| AOL Explorer | 0.01% |
| Apple Mail | 0.01% |
| BlackBerry Browser | 0.06% |
| Chrome | 23.88% |
| Chrome Mobile | 3.09% |
| Chromium | 0.02% |
| CometBird | 0.00% |
| Firefox | 8.07% |
| IBrowse | 0.00% |
| IceWeasel | 0.00% |
| IE | 5.92% |
| IE Mobile | 0.13% |

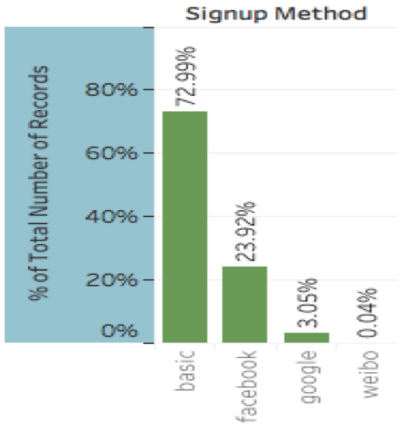
## Age

| Age | |
|---|---|
| Null | 46.50% |
| 1 | 0.00% |

## Gender



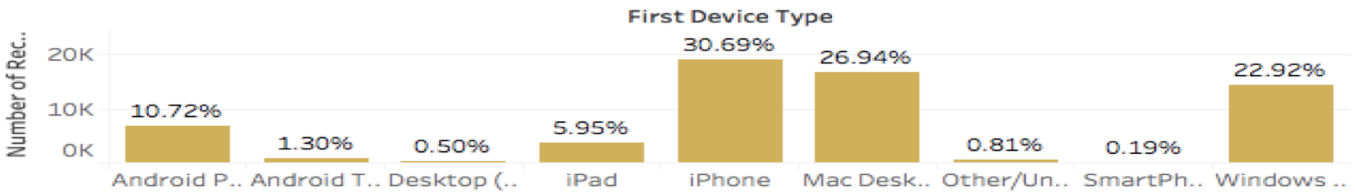

## Signup Method




## Date First Booking

| Date First B.. | |
|---|---|
| Null | 100.0% |

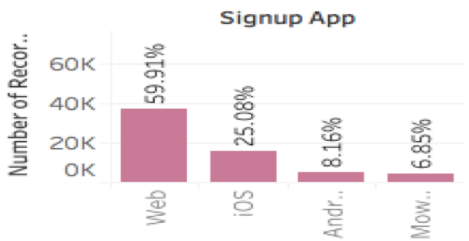
## First Device Type



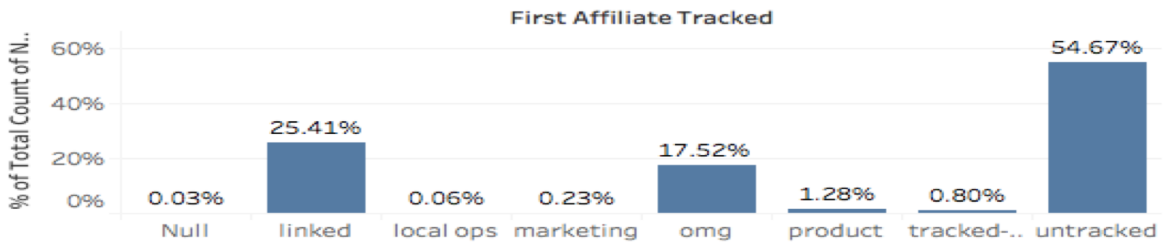

## Language

| Language | |
|---|---|
| -unknown- | 0.00% |
| ca | 0.00% |
| cs | 0.03% |
| da | 0.03% |
| de | 0.39% |
| el | 0.01% |
| en | 95.37% |
| es | 0.42% |
| fi | 0.01% |
| fr | 0.54% |
| hu | 0.01% |
| id | 0.00% |

## Signup App




## First Affiliated Tracked



# 3. Challenges with Dataset

- 46% Null values in age column of our dataset.
- Formatting problem in seconds elapsed column in Sessions dataset.
- Formatting problem with the date account created column in both train and test datasets.
- Formatting problem with timestamp first active column in both train and test datasets.
- Sessions dataset is too large (around 10M observations).
- 100% Null values in date first booking column in test dataset.

# 4. Data Cleaning

## ❑Null Values in Session Time

- *Created Dummy Variables in Sessions Dataset -> Applied linear model (lm) on time as target Variable -> Grouping time as per unique id (time_min/activity)*

Session Data - Before

| user_id | action | action_type | action_detail | device_type | secs_elapsed |
|---|---|---|---|---|---|
| d1mm9tcy42 | lookup | *NA* | *NA* | Windows Desktop | 319 |
| d1mm9tcy42 | search_results | click | view_search_results | Windows Desktop | 67753 |
| d1mm9tcy42 | lookup | *NA* | *NA* | Windows Desktop | 301 |
| d1mm9tcy42 | search_results | click | view_search_results | Windows Desktop | 22141 |
| d1mm9tcy42 | lookup | *NA* | *NA* | Windows Desktop | 435 |
| d1mm9tcy42 | search_results | click | view_search_results | Windows Desktop | 7703 |
| d1mm9tcy42 | lookup | *NA* | *NA* | Windows Desktop | 115 |
| d1mm9tcy42 | personalize | data | wishlist_content_update | Windows Desktop | 831 |
| d1mm9tcy42 | index | view | view_search_results | Windows Desktop | 20842 |
| d1mm9tcy42 | lookup | *NA* | *NA* | Windows Desktop | 683 |
| d1mm9tcy42 | search_results | click | view_search_results | Windows Desktop | 59274 |
| d1mm9tcy42 | lookup | *NA* | *NA* | Windows Desktop | 95 |
| d1mm9tcy42 | personalize | data | wishlist_content_update | Windows Desktop | 1399 |

*Size : 10567737 x 6 variables*

Session Data - After

| id | time_min |
|---|---|
| 00023iyk9l | 373.44178 |
| 0010k6l0om | 162.67388 |
| 001wyh0pz8 | 58.60749 |
| 0028jgx1x1 | 177.70185 |
| 002qnbzfs5 | 137.42681 |
| 0031awlkjq | 126.38951 |
| 0035hobuyj | 196.49041 |
| 00378ocvlh | 415.00506 |
| 00389675gq | 249.33382 |
| 003iamz20l | 166.72748 |

*Size : 135483 x 2 variables*

## ❑Cleaning Dates Format

```
> dac = as.data.frame(str_split_fixed(df_all$date_account_created, '-', 3))
> df_all['dac_year'] = dac[,1]
> df_all['dac_month'] = dac[,2]
> df_all['dac_day'] = dac[,3]
> df_all = df_all[,-c(which(colnames(df_all) %in% c('date_account_created')))]
>
>
> df_all[,'tfa_year'] = substring(as.character(df_all[,'timestamp_first_active']), 1, 4)
> df_all['tfa_month'] = substring(as.character(df_all['timestamp_first_active']), 5, 6)
> df_all['tfa_day'] = substring(as.character(df_all['timestamp_first_active']), 7, 8)
> df_all = df_all[,-c(which(colnames(df_all) %in% c('timestamp_first_active')))]
```

Date - Before

| date_account_created | timestamp_first_active |
|---|---|
| 2010-06-28 | 2.009032e+13 |
| 2011-05-25 | 2.009052e+13 |
| 2010-09-28 | 2.009061e+13 |
| 2011-12-05 | 2.009103e+13 |
| 2010-09-14 | 2.009121e+13 |
| 2010-01-01 | 2.010010e+13 |
| 2010-01-02 | 2.010010e+13 |
| 2010-01-03 | 2.010010e+13 |
| 2010-01-04 | 2.010010e+13 |

Date - After

| dac_year | dac_month | dac_day | tfa_year | tfa_month | tfa_day |
|---|---|---|---|---|---|
| 2014 | 05 | 14 | 2014 | 14 | 05 |
| 2013 | 12 | 03 | 2013 | 14 | 05 |
| 2013 | 02 | 27 | 2013 | 14 | 05 |
| 2013 | 07 | 08 | 2013 | 14 | 05 |
| 2011 | 12 | 27 | 2011 | 14 | 05 |
| 2012 | 09 | 21 | 2012 | 14 | 05 |
| 2012 | 04 | 25 | 2012 | 14 | 05 |

## ❑Null Values in Age

# ❑Null Values in Age

```
> df_all$age[df_all$age<=10 | df_all$age>=90]<- NA
> train_age <- subset(df_all, is.na(df_all$age) == F)
> test_age <- subset(df_all, is.na(df_all$age) == T)
> m_age <- lm(age ~ gender + affiliate_channel + signup_app + first_device_type, train_age)
> test_age$age <- predict(m_age,test_age)
> df_all <- rbind(train_age,test_age)
> df_all$age <- round(df_all$age, 0)
```

Age - Before

| id | date_account_created | timestamp_first_active | date_first_booking | gender | age |
|---|---|---|---|---|---|
| 5uwns89zht | 2014-07-01 | 2.01407e+13 | *NA* | FEMALE | 35 |
| jtl0dijy2j | 2014-07-01 | 2.01407e+13 | *NA* | -unknown- | *NA* |
| xx0ulgorjt | 2014-07-01 | 2.01407e+13 | *NA* | -unknown- | *NA* |
| 6c6puo6ix0 | 2014-07-01 | 2.01407e+13 | *NA* | -unknown- | *NA* |
| czqhjk3yfe | 2014-07-01 | 2.01407e+13 | *NA* | -unknown- | *NA* |
| szx28ujmhf | 2014-07-01 | 2.01407e+13 | *NA* | FEMALE | 28 |
| guenkfjcbq | 2014-07-01 | 2.01407e+13 | *NA* | MALE | 48 |
| tkpq0mlugk | 2014-07-01 | 2.01407e+13 | *NA* | -unknown- | *NA* |
| 3xtgd5p9dn | 2014-07-01 | 2.01407e+13 | *NA* | -unknown- | *NA* |

Age - After

| id | gender | age | s |
|---|---|---|---|
| 000wc9mlv3 | MALE | 40 | I |
| 001357912w | MALE | 38 | I |
| 001nvbxsvp | MALE | 41 | I |
| 001xf4efvm | FEMALE | 34 | I |
| 001y3jr7xc | FEMALE | 28 | I |
| 002dfbmaj5 | FEMALE | 26 | I |
| 002qnbzfs5 | FEMALE | 26 | I |
| 00389675gq | FEMALE | 26 | I |

- **Train Data - Before**

| id | date_account_created | timestamp_first_active | date_first_booking | gender | age | signup_method | signup_flow | language | affiliate_channel | affiliate_provider | first_affiliate_tracked | signup_app | first_device_type | first_browser | country_destination |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| gxn3p5htnn | 2010-06-28 | 2.009032e+13 | | -unknown- | NA | facebook | 0 | en | direct | direct | untracked | Web | Mac Desktop | Chrome | NDF |
| 820tgsjxq7 | 2011-05-25 | 2.009052e+13 | | MALE | 38 | facebook | 0 | en | seo | google | untracked | Web | Mac Desktop | Chrome | NDF |
| 4ft3gnwmtx | 2010-09-28 | 2.009061e+13 | 2010-08-02 | FEMALE | 56 | basic | 3 | en | direct | direct | untracked | Web | Windows Desktop | IE | US |
| bjjt8pjhuk | 2011-12-05 | 2.009103e+13 | 2012-09-08 | FEMALE | 42 | facebook | 0 | en | direct | direct | untracked | Web | Mac Desktop | Firefox | other |
| 87mebub9p4 | 2010-09-14 | 2.009121e+13 | 2010-02-18 | -unknown- | 41 | basic | 0 | en | direct | direct | untracked | Web | Mac Desktop | Chrome | US |
| osr2jwljor | 2010-01-01 | 2.010010e+13 | 2010-01-02 | -unknown- | NA | basic | 0 | en | other | other | omg | Web | Mac Desktop | Chrome | US |
| lsw9q7uk0j | 2010-01-02 | 2.010010e+13 | 2010-01-05 | FEMALE | 46 | basic | 0 | en | other | craigslist | untracked | Web | Mac Desktop | Safari | US |

*Size : 213451 x 16 variables*

- ✓ **Train Data - After**

| id | gender | age | signup_method | signup_flow | language | affiliate_channel | affiliate_provider | first_affiliate_tracked | signup_app | first_device_type | first_browser | time_min | dac_year | dac_month | dac_day | tfa_year | tfa_month | tfa_day |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 00023iyk9l | -unknown- | 31 | basic | 0 | en | direct | direct | omg | Web | Mac Desktop | Safari | 373.44178 | 2014 | 05 | 14 | 2014 | 14 | 05 |
| 000wc9mlv3 | MALE | 40 | basic | 0 | en | direct | direct | untracked | Web | Mac Desktop | Safari | 312.67270 | 2013 | 12 | 03 | 2013 | 14 | 05 |
| 001357912w | MALE | 38 | basic | 0 | en | direct | direct | untracked | Web | Windows Desktop | Firefox | 312.67270 | 2013 | 02 | 27 | 2013 | 14 | 05 |
| 001nvbxsvp | MALE | 41 | basic | 0 | en | direct | direct | untracked | Web | Mac Desktop | Firefox | 312.67270 | 2013 | 07 | 08 | 2013 | 14 | 05 |
| 001xf4efvm | FEMALE | 34 | basic | 2 | en | sem-non-brand | google | untracked | Web | Mac Desktop | Firefox | 312.67270 | 2011 | 12 | 27 | 2011 | 14 | 05 |
| 001y3jr7xc | FEMALE | 28 | basic | 3 | en | direct | direct | untracked | Web | Mac Desktop | Chrome | 312.67270 | 2012 | 09 | 21 | 2012 | 14 | 05 |

*Size : 213451 x 19 variables*

# 5. Modeling

# I. Random Forest

- Modeling Formula & Parameters:

```
> RF_M = randomForest(as.factor(country_destination) ~ gender + signup_method
+                     +signup_flow + language + affiliate_channel + affiliate_provider  + first_affiliate_tracked + signup_app
+                       +first_device_type + first_browser + timestamp_first_active,  importance = TRUE,
+                       ntrees = 10, data=train)
```

- Model Accuracy:

```
> confusionMatrix(predictions1, test$country_destination)
Confusion Matrix and Statistics

          Reference
Prediction    AU    CA    DE    ES    FR    GB    IT   NDF    NL other    PT    US
     AU        0     0     0     0     0     0     0     0     0     0     0     0
     CA        0     0     0     0     0     0     0     0     0     0     0     0
     DE        0     0     0     0     0     0     0     0     0     0     0     0
     ES        0     0     0     0     0     0     0     1     0     0     0     0
     FR        0     0     0     0     0     0     0     0     0     0     0     0
     GB        0     0     0     0     0     0     0     1     0     0     0     0
     IT        0     0     0     0     0     0     0     1     0     0     0     0
     NDF      57   189   126   292   637   279   342 21443    77  1312    26  7753
     NL        0     0     0     0     0     0     0     0     0     0     0     0
     other     0     0     0     0     0     0     0     3     1     0     0     2
     PT        0     0     0     0     0     0     0     0     0     0     0     0
     US       48   123    74   157   418   179   200  3583    61   694     9  4603

Overall Statistics

               Accuracy : 0.6101
                 95% CI : (0.6055, 0.6147)
    No Information Rate : 0.5864
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.195
 Mcnemar's Test P-Value : NA
```

Model Accuracy = 61.01%

# II. GBM Model

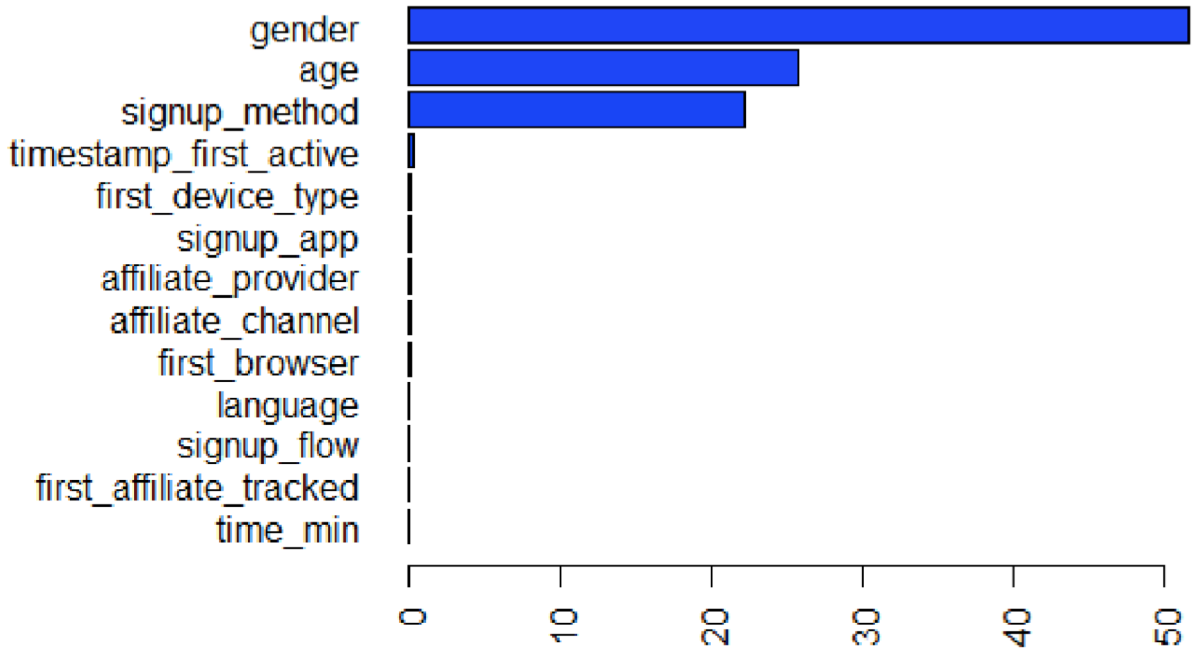
➢ Modeling Formula & Parameters:

```
> GBM_A = gbm(country_destination ~ gender + age + signup_method + signup_flow + language
+             + affiliate_channel + affiliate_provider + first_affiliate_tracked + signup_app
+             + first_device_type + first_browser + timestamp_first_active +time_min,
+             distribution = "multinomial", n.trees = 50, shrinkage = .1, interaction.depth =2,  data = train)
```

➢ Model Accuracy:

```
> confusionMatrix(predictions1, test$country_destination)
Confusion Matrix and Statistics

          Reference
Prediction    AU    CA    DE    ES    FR    GB    IT   NDF    NL other    PT    US
     AU        0     0     0     0     0     0     0     0     0     0     0     1
     CA        0     0     0     0     0     0     0     1     0     0     0     1
     DE        0     0     0     0     0     0     0     0     0     0     0     0
     ES        0     0     0     0     0     0     0     0     0     0     0     0
     FR        0     0     0     0     0     0     0     0     0     0     0     0
     GB        0     0     0     0     0     0     0     0     0     0     0     0
     IT        0     0     0     0     0     0     0     0     0     0     0     0
     NDF      72   206   140   311   741   331   404 22528    98  1477    28  8619
     NL        0     0     0     0     0     0     0     0     0     0     0     0
     other     0     0     0     0     1     0     0     0     0     0     0     1
     PT        0     0     0     0     0     0     0     0     0     0     0     0
     US       33   106    60   138   313   127   138  2503    41   529     7  3736

Overall Statistics

               Accuracy : 0.6152
                 95% CI : (0.6106, 0.6198)
    No Information Rate : 0.5864
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.1769
 Mcnemar's Test P-Value : NA
```

*Significance graph*




Model Accuracy = 61.52%

# III. XGBoost Model

- Modeling Formula & Parameters:

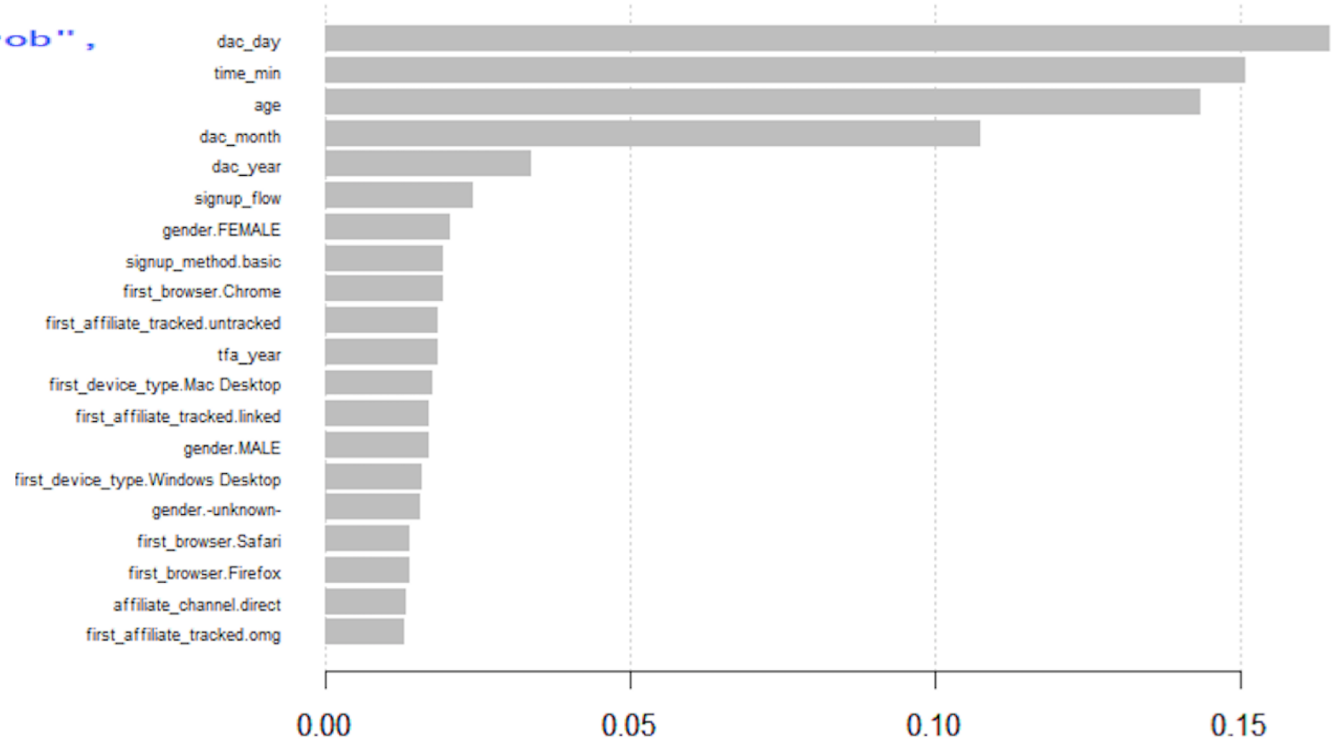
```
> xgb <- xgboost(data = data.matrix(X[,-1]),
+                label = y,
+                eta = 0.1,
+                max_depth = 20,
+                nround=250,
+                early_stopping_rounds = 100,
+                n_estimators = 100,
+                subsample = 0.5,
+                colsample_bytree = 0.5,
+                seed = 1,
+                eval_metric = "merror",
+                objective = "multi:softprob",
+                num_class = 12,
+                nthread = 3
+ )
```

- Model Accuracy:

```
> tail(e)
     iter train_merror
245   245      0.147280
246   246      0.146849
247   247      0.146127
248   248      0.145378
249   249      0.144989
250   250      0.144150
```

Model Accuracy = 1-0.1441 = 85.59%

*Significance graph*



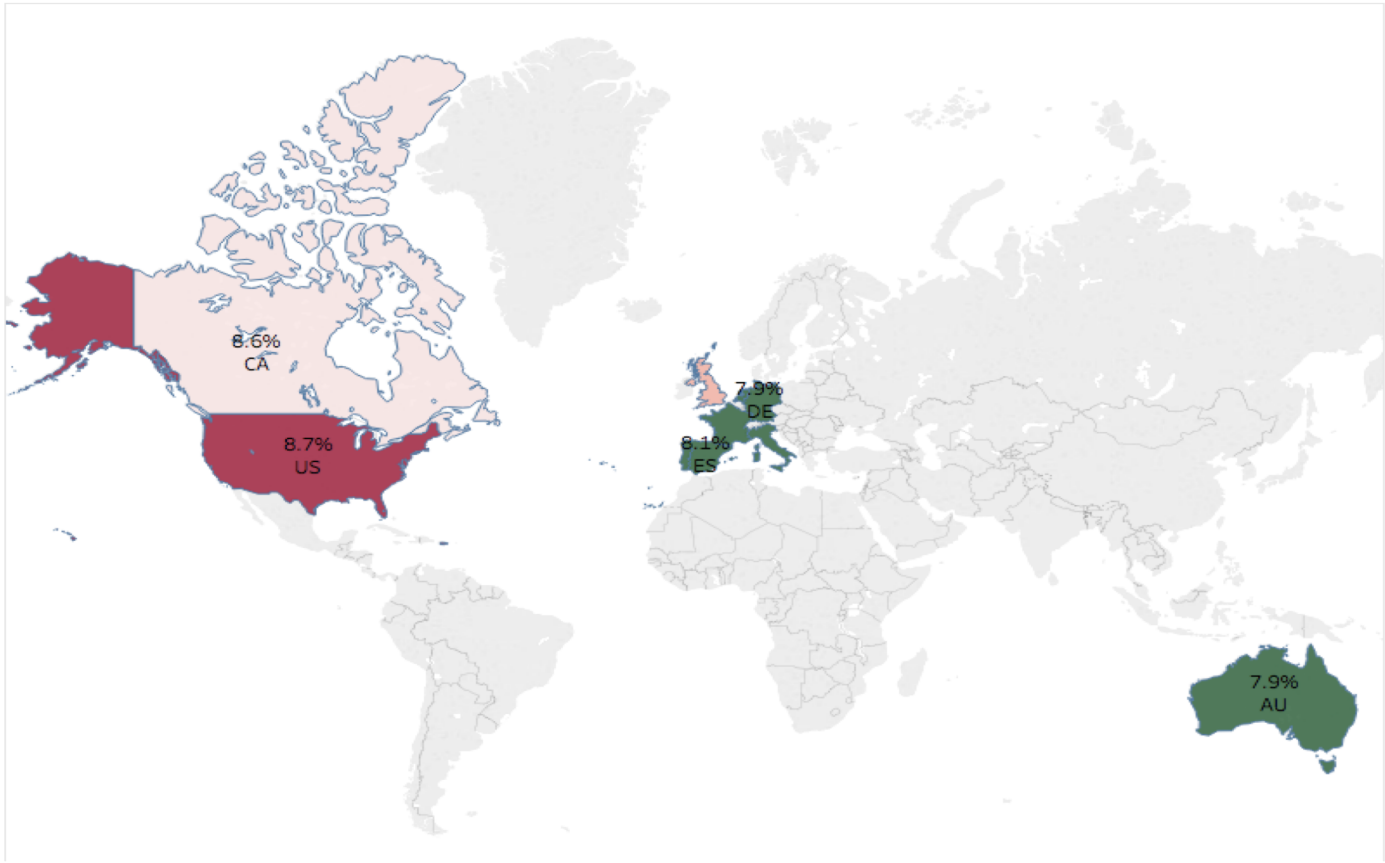
# 6. Predicting User Bookings

- Model Code:

```
> y_pred <- predict(xgb, data.matrix(X_test[,-1]))
```

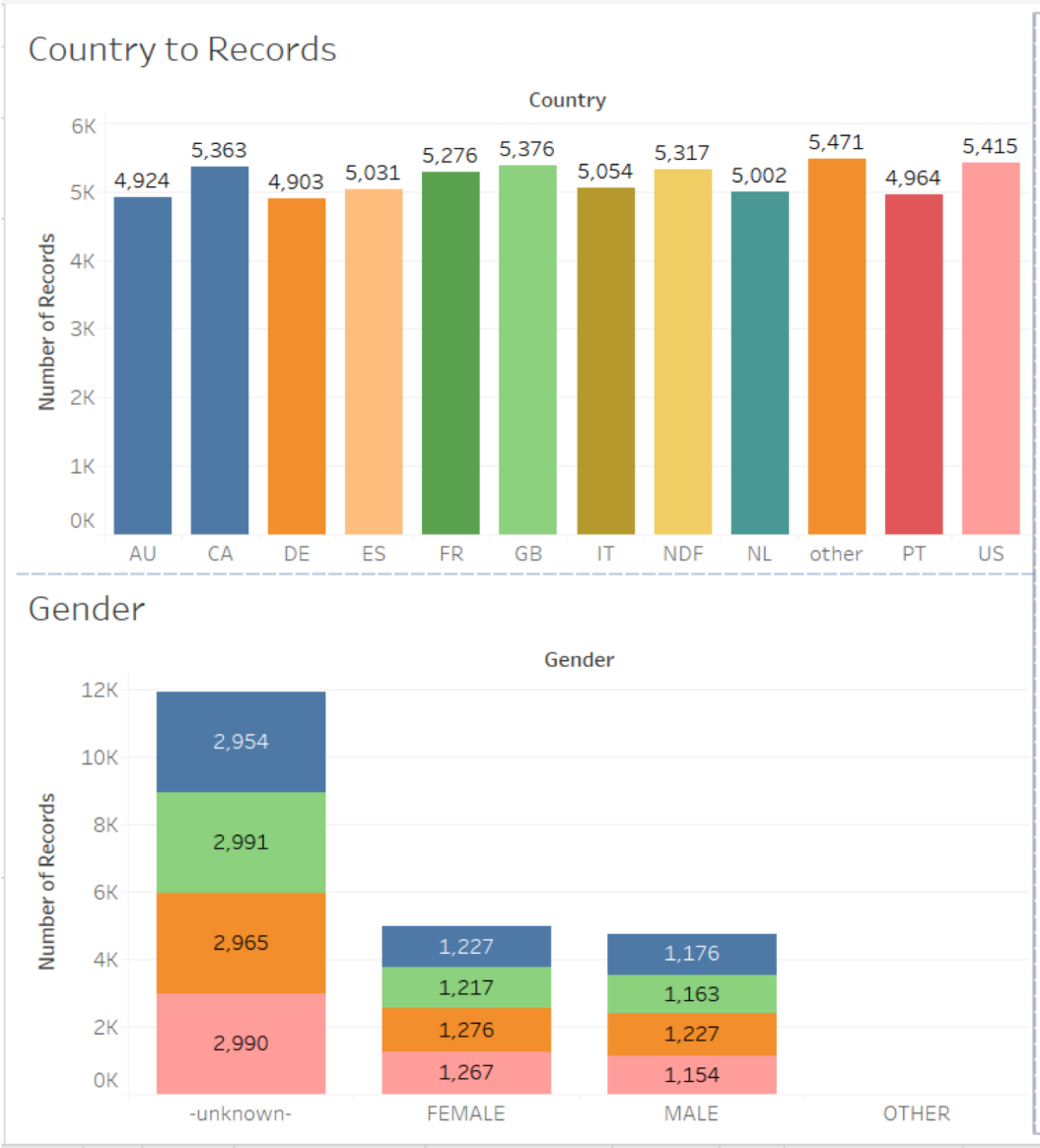
- Result:

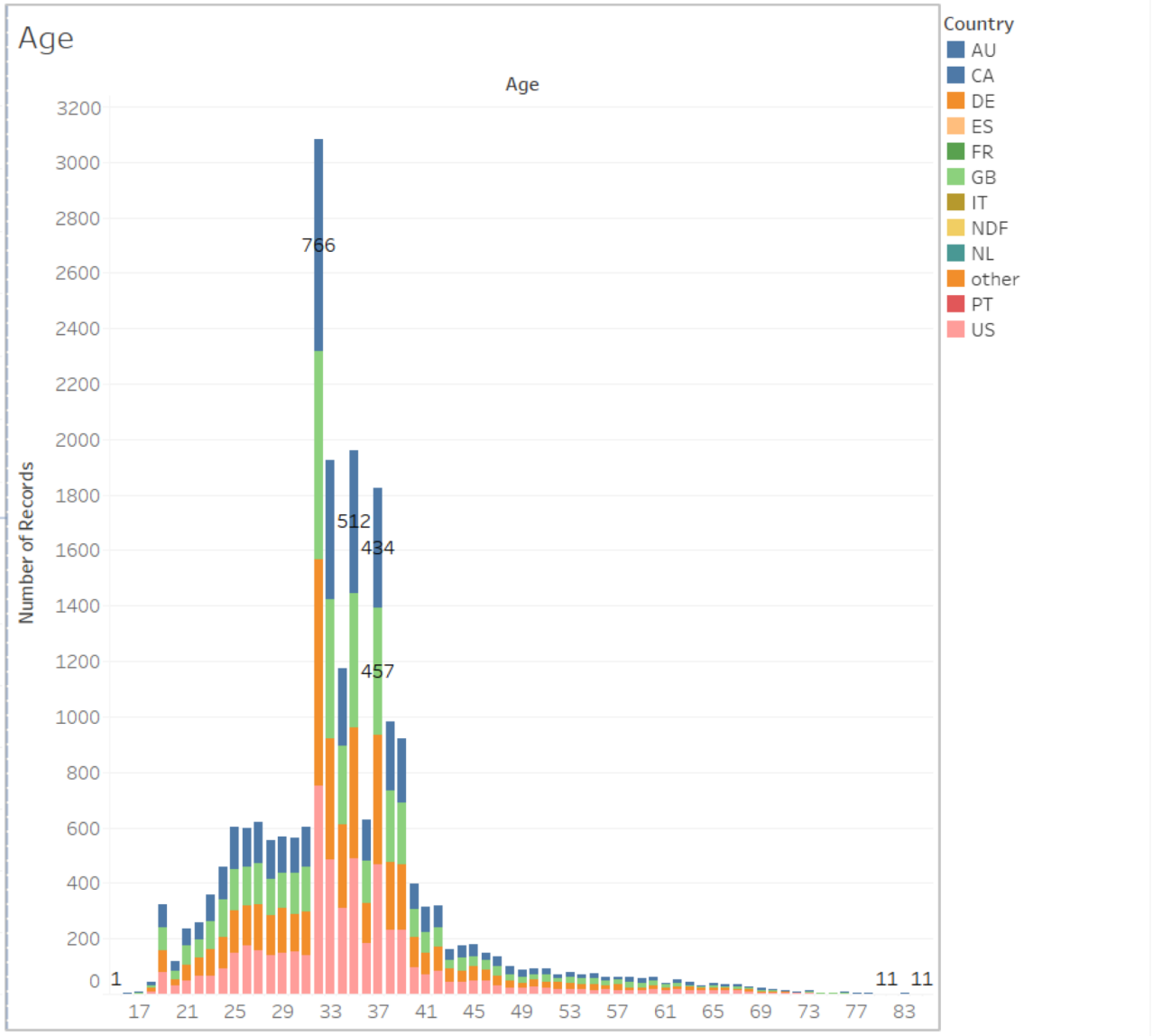
| | id | country |
|---|---|---|
| 1 | 0031awlkjq | NL |
| 2 | 00378ocvlh | FR |
| 3 | 0057snrdpu | ES |
| 4 | 0063bawn05 | AU |
| 5 | 0075z9e9xv | IT |
| 6 | 00an0o6c07 | DE |
| 7 | 00cdlcsaxu | FR |
| 8 | 00d41chbjd | NL |
| 9 | 00epe0uaxo | AU |
| 10 | 00ipxpr25h | other |
| 11 | 00j9fxfvut | CA |
| 12 | 00jcwpqbl2 | CA |
| 13 | 00kbbr6qom | FR |
| 14 | 00kpv7ok66 | NDF |
| 15 | 00rhdi533j | AU |



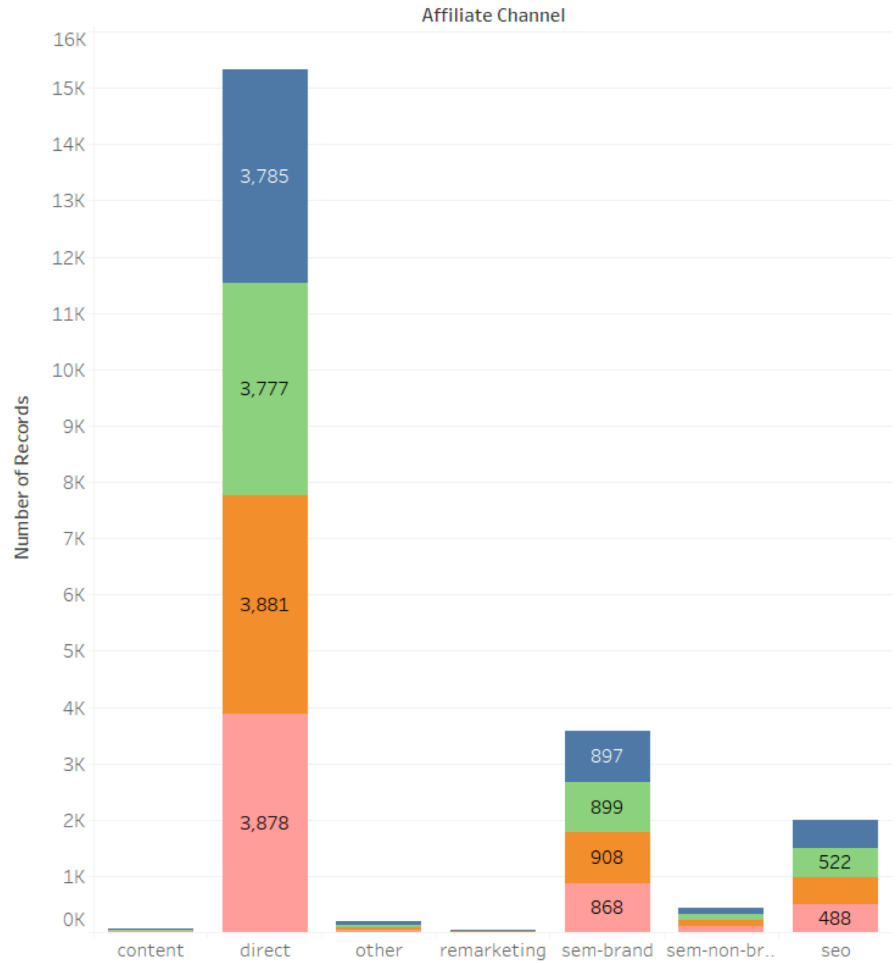

| Country | |
|---|---|
| AU | 7.9% |
| CA | 8.6% |
| DE | 7.9% |
| ES | 8.1% |
| FR | 8.5% |
| GB | 8.7% |
| IT | 8.1% |
| NDF | 8.6% |
| NL | 8.1% |
| other | 8.8% |
| PT | 8.0% |
| US | 8.7% |

# 7. Results

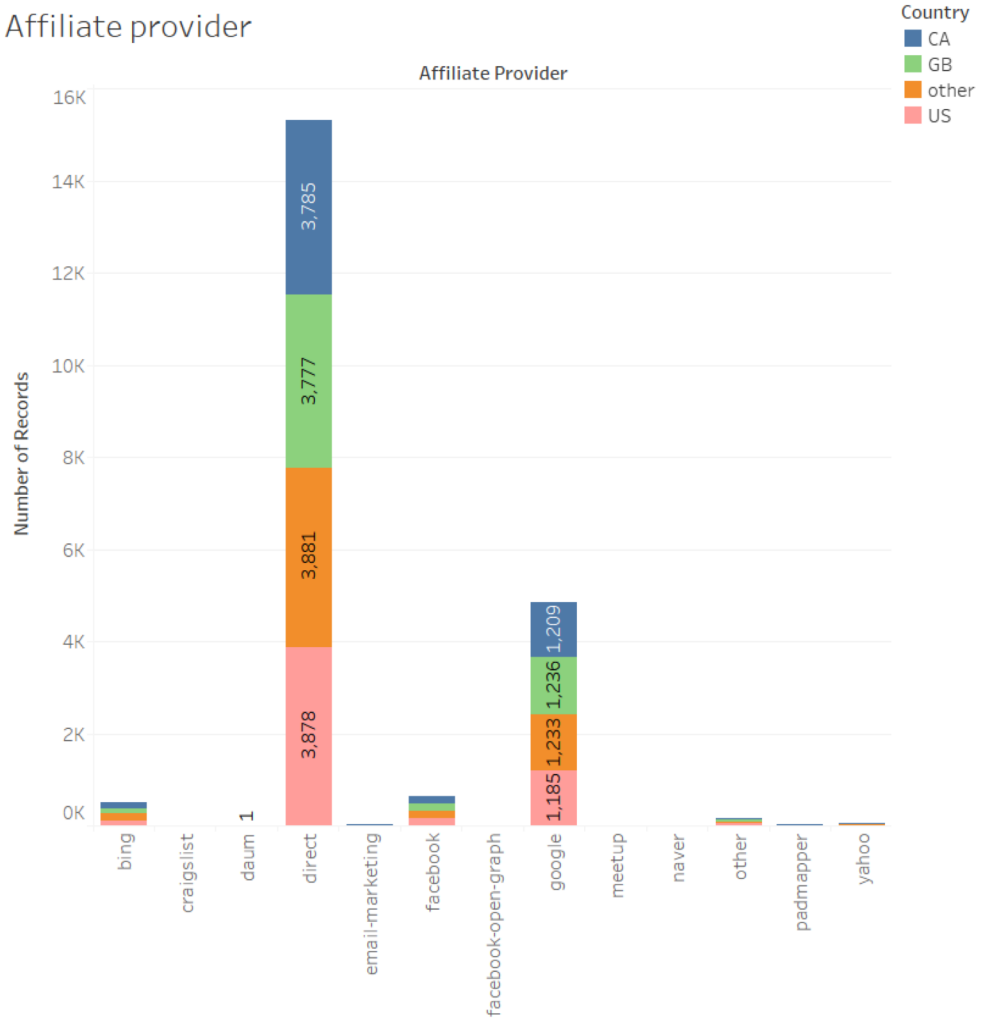

Country to Records
Country
Number of Records
6K
5K
4K
3K
2K
1K
0K
4,924
5,363
4,903
5,031
5,276
5,376
5,054
5,317
5,002
5,471
4,964
5,415
AU
CA
DE
ES
FR
GB
IT
NDF
NL
other
PT
US
Gender
Gender
Number of Records
12K
10K
8K
6K
4K
2K
0K
2,954
2,991
2,965
2,990
1,227
1,217
1,276
1,267
1,176
1,163
1,227
1,154
-unknown-
FEMALE
MALE
OTHER

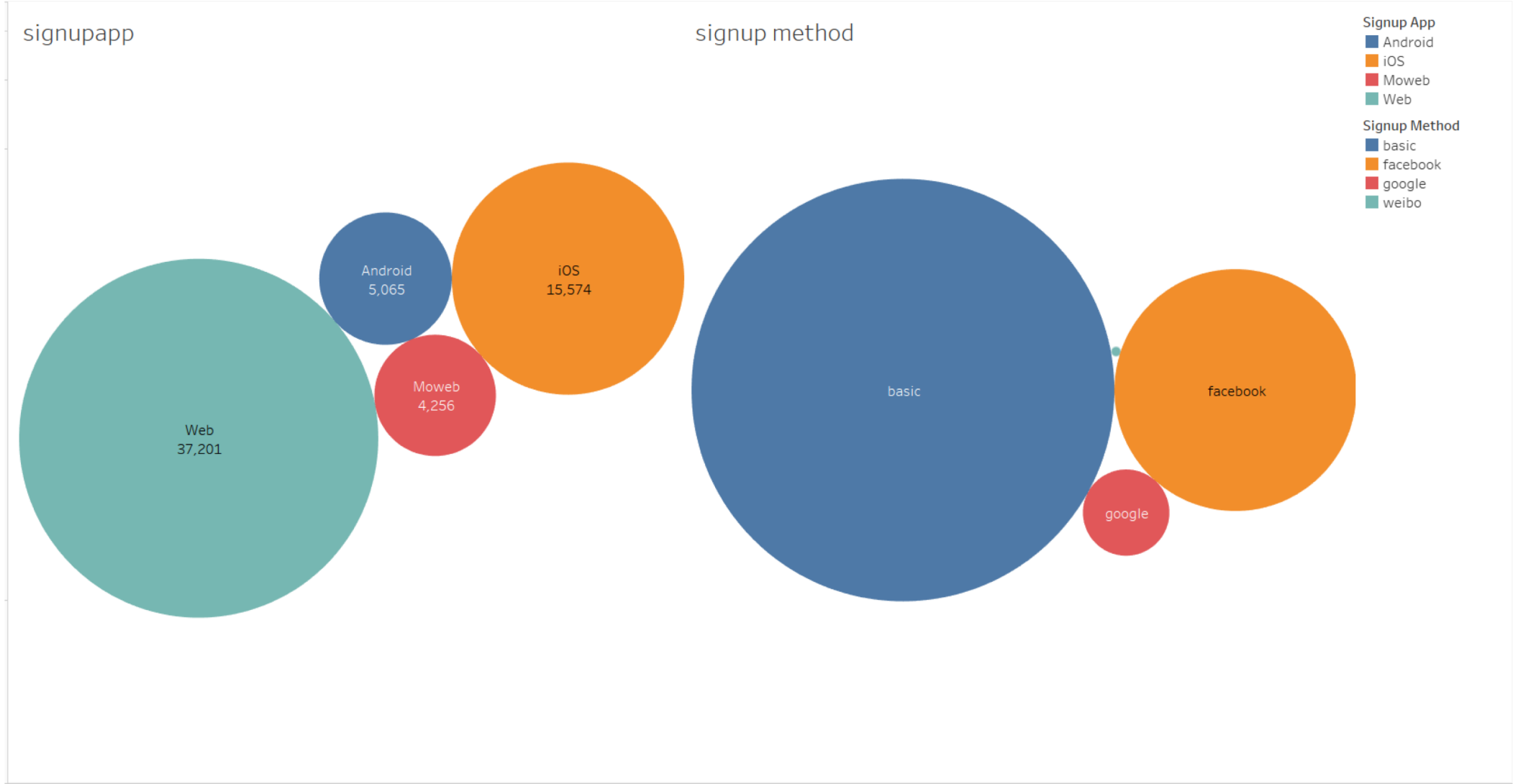

Age
Age
Number of Records
3200
3000
2800
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
766
512
434
457
1
11
11
17
21
25
29
33
37
41
45
49
53
57
61
65
69
73
77
83
Country
AU
CA
DE
ES
FR
GB
IT
NDF
NL
other
PT
US

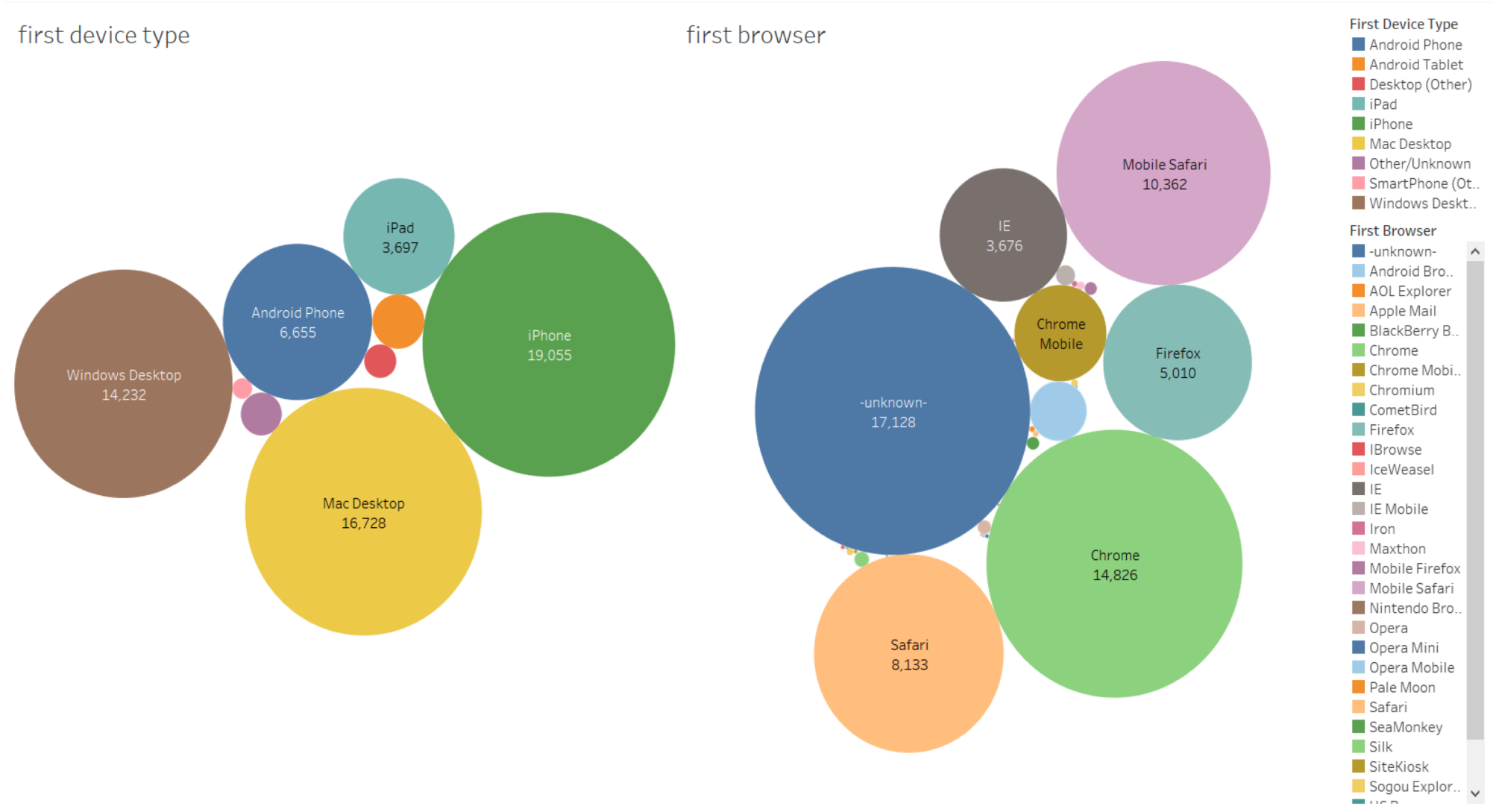
## Affliliate Channel

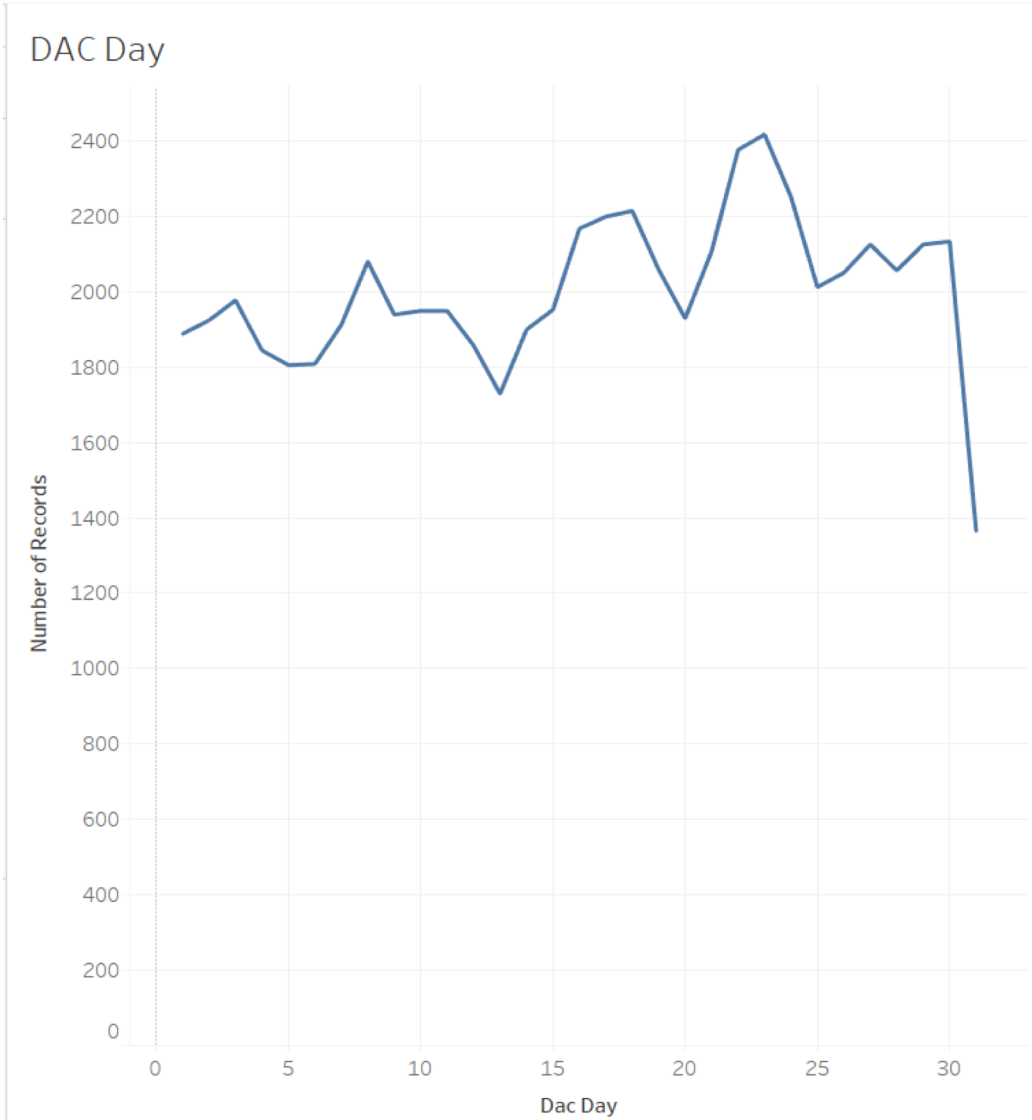


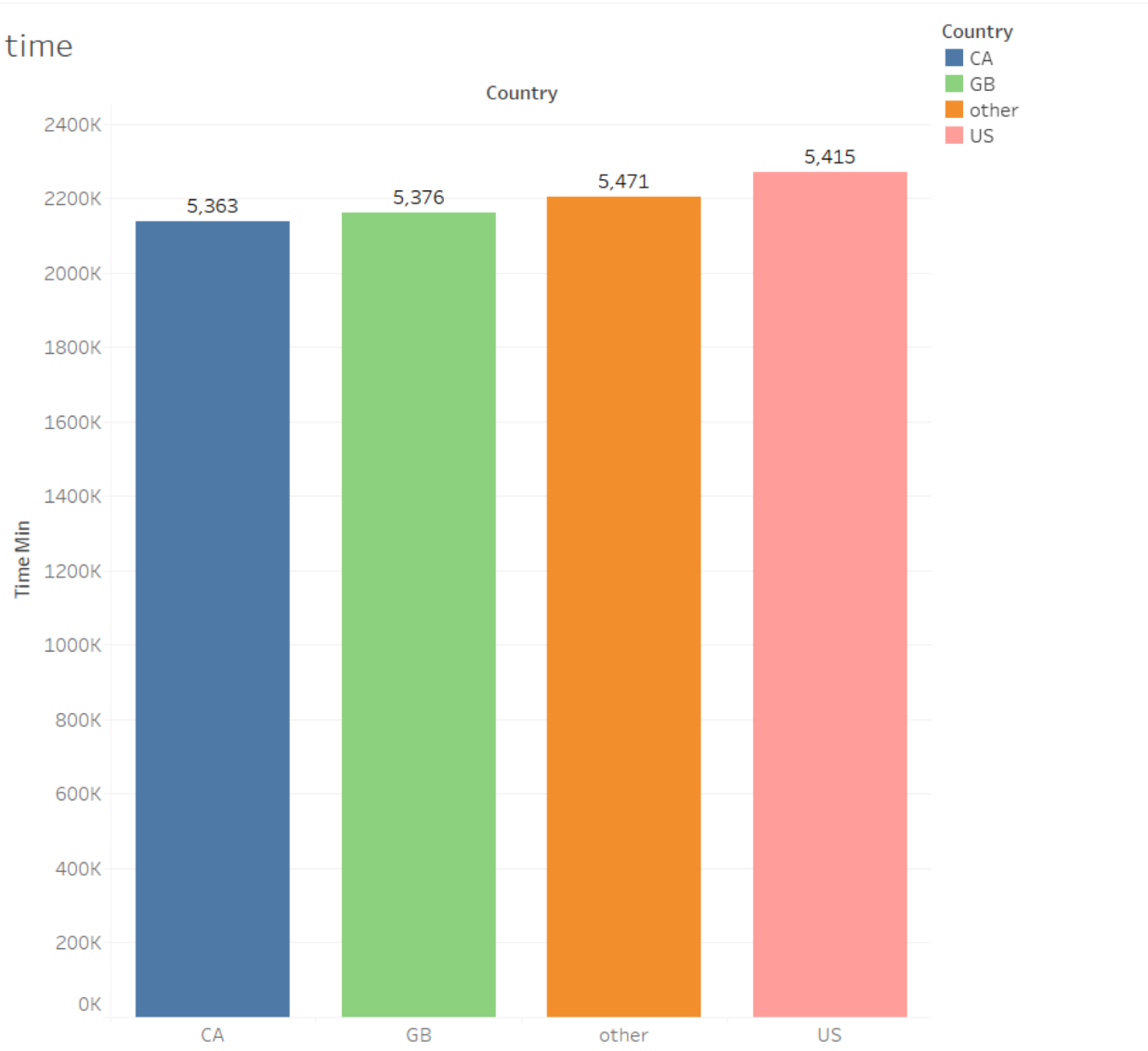

## Affiliate provider




signupapp
signup method
Signup App
Android
iOS
Moweb
Web
Signup Method
basic
facebook
google
weibo
Android
5,065
iOS
15,574
Moweb
4,256
Web
37,201
basic
facebook
google


first device type
iPad
3,697
Android Phone
6,655
Windows Desktop
14,232
iPhone
19,055
Mac Desktop
16,728
first browser
Mobile Safari
10,362
IE
3,676
Chrome
Mobile
Firefox
5,010
-unknown-
17,128
Chrome
14,826
Safari
8,133
First Device Type
Android Phone
Android Tablet
Desktop (Other)
iPad
iPhone
Mac Desktop
Other/Unknown
SmartPhone (Ot..
Windows Deskt..
First Browser
-unknown-
Android Bro..
AOL Explorer
Apple Mail
BlackBerry B..
Chrome
Chrome Mobi..
Chromium
CometBird
Firefox
IBrowse
IceWeasel
IE
IE Mobile
Iron
Maxthon
Mobile Firefox
Mobile Safari
Nintendo Bro..
Opera
Opera Mini
Opera Mobile
Pale Moon
Safari
SeaMonkey
Silk
SiteKiosk
Sogou Explor..


DAC Day
Number of Records
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
5
10
15
20
25
30
Dac Day


time
Country
CA
GB
other
US
Country
Time Min
2400K
2200K
2000K
1800K
1600K
1400K
1200K
1000K
800K
600K
400K
200K
0K
5,363
5,376
5,471
5,415
CA
GB
other
US

# Thank You!