**Train.users.csv**

1. Totally 213,452 rows with 16 columns, with substantial missing values in column Age

> str(train.users.data)

'data.frame': 213451 obs. of 16 variables:

$ id : chr "gxn3p5htnn" "820tgsjxq7" "4ft3gnwmtx" "bjjt8pjhuk" ...

$ date\_account\_created : chr "6/28/2010" "5/25/2011" "9/28/2010" "12/5/2011" ...

$ timestamp\_first\_active : num 2.01e+13 2.01e+13 2.01e+13 2.01e+13 2.01e+13 ...

$ date\_first\_booking : chr "" "" "8/2/2010" "9/8/2012" ...

$ gender : chr "-unknown-" "MALE" "FEMALE" "FEMALE" ...

$ age : int NA 38 56 42 41 NA 46 47 50 46 ...

$ signup\_method : chr "facebook" "facebook" "basic" "facebook" ...

$ signup\_flow : int 0 0 3 0 0 0 0 0 0 0 ...

$ language : chr "en" "en" "en" "en" ...

$ affiliate\_channel : chr "direct" "seo" "direct" "direct" ...

$ affiliate\_provider : chr "direct" "google" "direct" "direct" ...

$ first\_affiliate\_tracked: chr "untracked" "untracked" "untracked" "untracked" ...

$ signup\_app : chr "Web" "Web" "Web" "Web" ...

$ first\_device\_type : chr "Mac Desktop" "Mac Desktop" "Windows Desktop" "Mac Desktop" ...

$ first\_browser : chr "Chrome" "Chrome" "IE" "Firefox" ...

$ country\_destination : chr "NDF" "NDF" "US" "other" ...

1. Extract Day, Weekday, Month, Year from columns – Date\_Account\_Created and Date\_first\_booking

This will help show the impact on bookings based on month, holidays etc like most popular dates for booking

1. Age Data Cleanup – There are many values in thousands and some in single digits. Assume an age range – 15 to 100, and then assume the 4 digit years from 1924-1995 as birth years.

Any age outside the valid range is set as NaN

87,990 users have missing age ie 41% of users. We have to see if there is any value in identifying users who have not provided age before we remove those observations.

20,376 users with missing age have made a booking (excluding NDF) ie 23% of users (20,376/87,990\*100)

123,088 users have valid age ie 57% of users.

67,213 users with valid age have made a booking(excluding NDF) ie 55% of users

(67,213/123088 \*100)

This suggests there is value in keeping the observations with missing age.

1. Country-Destination has 12 valid classes with NDF meaning no booking.

About 123,489 users ie 58% of users have not made a booking.

|  |  |
| --- | --- |
| NDF | 123,489 |
| AU | 532 |
| CA | 1,406 |
| DE | 1,044 |
| ES | 2,219 |
| FR | 4,962 |
| GB | 2,287 |
| IT | 2,787 |
| NL | 750 |
| other | 9,931 |
| PT | 215 |
| US | 61,457 |

87,590 users have made a booking, out of that 61,457 users have chosen US –

That is among those that book, 70% chose to do it in US.

Further analysis shows that the bulk of booking is done by users in age-group 30-60.

The bookings drop significantly once the user’s age exceeds 60.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Age group** | **GB count** | **GB %** | **US count** | **US %** | **FR count** | **FR %** | **ES count** | **ES %** | **IT count** | **IT %** | **Other count** | **Other %** |
| 15-30 | 563 | 32 | 17,514 | 37 | 1,185 | 32 | 685 | 41 | 697 | 35 | 2,598 | 35 |
| 30-60 | 1,049 | 60 | 28,124 | 59 | 2,256 | 62 | 915 | 55 | 1,168 | 59 | 4,459 | 60 |
| >60 | 126 | 7 | 2,028 | 4 | 210 | 6 | 76 | 5 | 123 | 6 | 355 | 5 |
| **Total** | **1,738** |  | **47,666** |  | **3,651** |  | **1,676** |  | **1,988** |  | **7,412** |  |

1. There are 29,018 users whose gender is “unknown”. This is less than bookings by users with known gender (Females – 31,993 and Male – 27,721). Others book less frequently.
2. Users with the 'google' signup\_method book less frequently than 'basic' or 'facebook'
3. Users with signup\_flow 3 book more frequently than any other category (5615 users)
4. Language - there are a large number of languages represented even though major bookings were done by users in US
5. Users with affiliate\_channel 'content' book less frequently than other categories
6. Users with affiliate\_provider 'craigslist', direct', and 'google' book more frequently than other categories. We don’t know why the google affiliate channel is more effective than the google sign up method
7. Users with first\_affiliate\_tracked 'local ops' book less frequently than other categories
8. Users with signup\_app 'Web' booked the most frequently, while those with 'Android' booked the least
9. Users with first\_device\_type 'Mac\_Desktop' booked the most frequently, while those with 'Android Phone' booked the least. This could mean the app on Android is not user friendly

**Sessions.data**

1. 10 M rows with multiple rows per user-id.
2. 3563 rows don’t have valid user-ids. Could be ignored.
3. Fill blanks as NULL.
4. 40,000 rows of the train data have no corresponding entry in sessions.csv. The same is only true of 265 rows in the test data

> str(sessions.data)

'data.frame': 10567737 obs. of 6 variables:

$ user\_id : chr "d1mm9tcy42" "d1mm9tcy42" "d1mm9tcy42" "d1mm9tcy42" ...

$ action : chr "lookup" "search\_results" "lookup" "search\_results" ...

$ action\_type : chr "" "click" "" "click" ...

$ action\_detail: chr "" "view\_search\_results" "" "view\_search\_results" ...

$ device\_type : chr "Windows Desktop" "Windows Desktop" "Windows Desktop" "Windows Desktop" ...

$ secs\_elapsed : num 319 67753 301 22141 435 ...

df=as.data.frame(sessions.data%>%group\_by(action)%>%summarise(count=n())%>%arrange(desc(count)))

**360 unique action items – Top 50 of them with frequency count**

1 show 2768278

2 index 843699

3 search\_results 725226

4 personalize 706824

5 search 536057

6 ajax\_refresh\_subtotal 487744

7 update 365130

8 similar\_listings 364624

9 social\_connections 339000

10 reviews 320591

11 active 188036

12 similar\_listings\_v2 168788

13 lookup 162041

14 create 155887

15 dashboard 152952

16 header\_userpic 141830

17 collections 124417

18 edit 109083

19 campaigns 105028

20 track\_page\_view 81117

21 79626

22 unavailabilities 78317

23 qt2 64651

24 notifications 59392

25 confirm\_email 58726

26 requested 57034

27 identity 53631

28 ajax\_check\_dates 52517

29 show\_personalize 50434

30 ask\_question 44063

31 listings 43656

32 authenticate 42930

33 calendar\_tab\_inner2 42198

34 travel\_plans\_current 38195

35 edit\_verification 33434

36 ajax\_lwlb\_contact 33413

37 other\_hosting\_reviews\_first 31045

38 recommendations 30592

39 impressions 27787

40 manage\_listing 26618

41 click 25571

42 complete\_status 24674

43 ajax\_photo\_widget\_form\_iframe 24442

44 payment\_instruments 23573

45 message\_to\_host\_focus 21636

46 verify 21372

47 payment\_methods 19800

48 cancellation\_policies 19262

49 callback 19132

50 settings 16766

> df=as.data.frame(sessions.data%>%group\_by(action\_type)%>%summarise(count=n())%>%arrange(desc(count)))

**10 unique action\_type items**

action\_type count

1 view 3560902

2 data 2103770

3 click 1996183

4 1126204

5 -unknown- 1031170

6 submit 623357

7 message\_post 87103

8 partner\_callback 19132

9 booking\_request 18773

10 modify 1139

11 booking\_response 4

> df=as.data.frame(sessions.data%>%group\_by(action\_detail)%>%summarise(count=n())%>%arrange(desc(count)))

**156 unique items**

action\_detail count

1 view\_search\_results 1776885

2 p3 1376550

3 1126204

4 -unknown- 1031141

5 wishlist\_content\_update 706824

6 user\_profile 656839

7 change\_trip\_characteristics 487744

8 similar\_listings 364624

9 user\_social\_connections 336799

10 update\_listing 269779

11 listing\_reviews 269021

12 dashboard 152952

13 user\_wishlists 152672

14 header\_userpic 141830

15 message\_thread 132395

16 edit\_profile 108686

17 message\_post 87103

18 contact\_host 81041

19 unavailable\_dates 77795

20 confirm\_email\_link 58422

21 create\_user 54878

22 change\_contact\_host\_dates 52517

23 user\_profile\_content\_update 50434

24 user\_reviews 49665

25 p5 48167

26 login 38378

27 your\_trips 38195

28 p1 37976

29 notifications 37355

30 profile\_verifications 33434

31 reservations 32481

32 user\_listings 31971

33 your\_listings 30980

34 listing\_recommendations 30215

35 update\_user 29008

36 create\_phone\_numbers 28336

37 p4 28030

38 update\_listing\_description 26980

39 update\_user\_profile 26689

40 manage\_listing 26618

41 payment\_instruments 23483

42 account\_notification\_settings 21802

43 message\_to\_host\_focus 21636

44 signup 21494

45 cancellation\_policies 19262

46 oauth\_response 19132

47 message\_inbox 18718

48 view\_listing 18474

49 message\_to\_host\_change 14224

50 list\_your\_space 14098

> df=as.data.frame(sessions.data%>%group\_by(device\_type)%>%summarise(count=n())%>%arrange(desc(count)))

>df

14 unique device types with frequency count

device\_type count

1 Mac Desktop 3594286

2 Windows Desktop 2658539

3 iPhone 2105031

4 Android Phone 839637

5 iPad Tablet 683414

6 Android App Unknown Phone/Tablet 273652

7 -unknown- 211279

8 Tablet 139886

9 Linux Desktop 28373

10 Chromebook 22348

11 iPodtouch 8198

12 Windows Phone 2047

13 Blackberry 979

14 Opera Phone 68

**Comparison of categorical variables in Training user’s data Vs Test user’s data**

Train user’s data – 213,452 unique ids and Test user’s data – 62,097 unique ids

**The categorical variables are**

**Categorical variables**

**1. Gender**

> unique(train.users.data$gender)

[1] -unknown- MALE FEMALE OTHER

Levels: -unknown- FEMALE MALE OTHER

> unique(test.users.data$gender)

[1] FEMALE -unknown- MALE OTHER

Levels: -unknown- FEMALE MALE OTHER

**2. Signup\_method**

> unique(train.users.data$signup\_method)

[1] facebook basic google

Levels: basic facebook google

> unique(test.users.data$signup\_method)

[1] facebook basic google weibo

Levels: basic facebook google weibo

**3. Language**

> unique(train.users.data$language)

25 Levels: ca cs da de el en es fi fr hr hu id is it ja ko nl no pl pt ru ... zh

> unique(test.users.data$language)

24 Levels: -unknown- ca cs da de el en es fi fr hu id it ja ko nl no pl pt ... zh

**4. Affiliate\_channel**

> unique(train.users.data$affiliate\_channel)

[1] direct seo other sem-non-brand content

[6] sem-brand remarketing api

Levels: api content direct other remarketing sem-brand sem-non-brand seo

> unique(test.users.data$affiliate\_channel)

[1] direct sem-brand sem-non-brand seo remarketing

[6] other content

Levels: content direct other remarketing sem-brand sem-non-brand seo

**5. Affiliate\_provider**

> unique(train.users.data$affiliate\_provider)

[1] direct google other

[4] craigslist facebook vast

[7] bing meetup facebook-open-graph

[10] email-marketing yahoo padmapper

[13] gsp wayn naver

[16] baidu yandex daum

18 Levels: baidu bing craigslist daum direct email-marketing ... yandex

> unique(test.users.data$affiliate\_provider)

[1] direct google bing

[4] facebook other craigslist

[7] padmapper email-marketing yahoo

[10] baidu naver gsp

[13] facebook-open-graph meetup vast

[16] daum yandex

1. Levels: baidu bing craigslist daum direct email-marketing ... yandex

**6. First\_affiliate\_tracked**

> unique(train.users.data$first\_affiliate\_tracked)

[1] untracked omg linked tracked-other

[6] product marketing local ops

Levels: linked local ops marketing omg product tracked-other untracked

> unique(test.users.data$first\_affiliate\_tracked)

[1] untracked linked omg product marketing

[6] tracked-other local ops

Levels: linked local ops marketing omg product tracked-other untracked

1. **Signup\_app**

> unique(train.users.data$signup\_app)

[1] Web Moweb iOS Android

Levels: Android iOS Moweb Web

> unique(test.users.data$signup\_app)

[1] Moweb Web iOS Android

Levels: Android iOS Moweb Web

1. **first\_device\_type**

> unique(train.users.data$first\_device\_type)

[1] Mac Desktop Windows Desktop iPhone Other/Unknown

[5] Desktop (Other) Android Tablet iPad Android Phone

[9] SmartPhone (Other)

9 Levels: Android Phone Android Tablet Desktop (Other) iPad ... Windows Desktop

> unique(test.users.data$first\_device\_type)

[1] iPhone Windows Desktop Mac Desktop iPad

[5] Android Tablet Android Phone Desktop (Other) Other/Unknown

[9] SmartPhone (Other)

9 Levels: Android Phone Android Tablet Desktop (Other) iPad ... Windows Desktop

1. **first\_browser**

> unique(train.users.data$first\_browser)

[1] Chrome IE Firefox

[4] Safari -unknown- Mobile Safari

[7] Chrome Mobile RockMelt Chromium

[10] Android Browser AOL Explorer Palm Pre web browser

[13] Mobile Firefox Opera TenFourFox

[16] IE Mobile Apple Mail Silk

[19] Camino Arora BlackBerry Browser

[22] SeaMonkey Iron Sogou Explorer

[25] IceWeasel Opera Mini SiteKiosk

[28] Maxthon Kindle Browser CoolNovo

[31] Conkeror wOSBrowser Google Earth

[34] Crazy Browser Mozilla OmniWeb

[37] PS Vita browser NetNewsWire CometBird

[40] Comodo Dragon Flock Pale Moon

[43] Avant Browser Opera Mobile Yandex.Browser

[46] TheWorld Browser SlimBrowser Epic

[49] Stainless Googlebot Outlook 2007

[52] IceDragon

52 Levels: -unknown- Android Browser AOL Explorer Apple Mail ... Yandex.Browser

> unique(test.users.data$first\_browser)

[1] Mobile Safari Chrome IE Safari

[5] -unknown- Firefox Chrome Mobile Android Browser

[9] IE Mobile BlackBerry Browser Opera Silk

[13] Mobile Firefox AOL Explorer SeaMonkey Opera Mobile

[17] wOSBrowser Chromium Apple Mail Maxthon

[21] IBrowse Sogou Explorer Iron Yandex.Browser

[25] SiteKiosk Pale Moon Nintendo Browser Opera Mini

[29] CometBird IceWeasel UC Browser

31 Levels: -unknown- Android Browser AOL Explorer ... Yandex.Browser

**24 new browsers in Train users, 3 new ones in Test users data.**

|  |  |
| --- | --- |
| Android Browser | Android Browser |
| AOL Explorer | AOL Explorer |
| Apple Mail | Apple Mail |
| Arora |  |
| Avant Browser |  |
| BlackBerry Browser | BlackBerry Browser |
| Camino |  |
| Chrome | Chrome |
| Chrome Mobile | Chrome Mobile |
| Chromium | Chromium |
| CometBird | CometBird |
| Comodo Dragon |  |
| Conkeror |  |
| CoolNovo |  |
| Crazy Browser |  |
| Epic |  |
| Firefox | Firefox |
| first\_browser | first\_browser |
| Flock |  |
| Google Earth |  |
| Googlebot |  |
| IceDragon |  |
| IceWeasel | IceWeasel |
| IE | IE |
| IE Mobile | IE Mobile |
| Iron | Iron |
| Kindle Browser |  |
| Maxthon | Maxthon |
| Mobile Firefox | Mobile Firefox |
| Mobile Safari | Mobile Safari |
| Mozilla |  |
| NetNewsWire |  |
| OmniWeb |  |
| Opera | Opera |
| Opera Mini | Opera Mini |
| Opera Mobile | Opera Mobile |
| Outlook 2007 |  |
| Pale Moon | Pale Moon |
| Palm Pre web browser |  |
| PS Vita browser |  |
| RockMelt |  |
| Safari | Safari |
| SeaMonkey | SeaMonkey |
| Silk | Silk |
| SiteKiosk | SiteKiosk |
| SlimBrowser |  |
| Sogou Explorer | Sogou Explorer |
| Stainless |  |
| TenFourFox |  |
| TheWorld Browser |  |
| -unknown- | -unknown- |
| wOSBrowser | wOSBrowser |
| Yandex.Browser | Yandex.Browser |
|  | UC Browser |
|  | IBrowse |
|  | Nintendo Browser |

1. **country\_destination**

> unique(train.users.data$country\_destination)

[1] NDF US other FR CA GB ES IT PT NL DE AU

Levels: AU CA DE ES FR GB IT NDF NL other PT US

> unique(test.users.data$country\_destination)

NULL