# The Airbnb data

This Dataset contains Airbnb data collected online. It contains infomation about when and through which channel users signed up and patronized their service.

In [1]:
```
#importing the necessary libraries
import numpy as np
import pandas as pd
import datetime as dt
import matplotlib.pyplot as plt
import warnings
warnings.filterwarnings("ignore")
```

In [2]:
```
#importing the file
airbnbdata=pd.read_csv('airbnb_data.csv')

airbnbdata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 213451 entries, 0 to 213450
Data columns (total 16 columns):
 #   Column                   Non-Null Count   Dtype  
---  ------                   --------------   -----  
 0   id                       213451 non-null  object 
 1   date_account_created     213451 non-null  object 
 2   timestamp_first_active   213451 non-null  int64  
 3   date_first_booking       88908 non-null   object 
 4   gender                   213451 non-null  object 
 5   age                      125461 non-null  float64
 6   signup_method            213451 non-null  object 
 7   signup_flow              213451 non-null  int64  
 8   language                 213451 non-null  object 
 9   affiliate_channel        213451 non-null  object 
 10  affiliate_provider       213451 non-null  object 
 11  first_affiliate_tracked  207386 non-null  object 
 12  signup_app               213451 non-null  object 
 13  first_device_type        213451 non-null  object 
 14  first_browser            213451 non-null  object 
 15  country_destination      213451 non-null  object 
dtypes: float64(1), int64(2), object(13)
memory usage: 26.1+ MB
```

In [3]:
```
#taking a peep
airbnbdata.head(20)
```

Out[3]:

| | id | date_account_created | timestamp_first_active | date_first_booking | gender | age | sign |
|---|---|---|---|---|---|---|---|
| **0** | gxn3p5htnn | 2010-06-28 | 20090319043255 | NaN | -unknown- | NaN | |
| **1** | 820tgsjxq7 | 2011-05-25 | 20090523174809 | NaN | MALE | 38.0 | |
| **2** | 4ft3gnwmtx | 2010-09-28 | 20090609231247 | 2010-08-02 | FEMALE | 56.0 | |
| **3** | bjjt8pjhuk | 2011-12-05 | 20091031060129 | 2012-09-08 | FEMALE | 42.0 | |

| | id | date_account_created | timestamp_first_active | date_first_booking | gender | age | signu |
|---|---|---|---|---|---|---|---|
| **4** | 87mebub9p4 | 2010-09-14 | 20091208061105 | 2010-02-18 | -unknown- | 41.0 | |
| **5** | osr2jwljor | 2010-01-01 | 20100101215619 | 2010-01-02 | -unknown- | NaN | |
| **6** | lsw9q7uk0j | 2010-01-02 | 20100102012558 | 2010-01-05 | FEMALE | 46.0 | |
| **7** | 0d01nltbrs | 2010-01-03 | 20100103191905 | 2010-01-13 | FEMALE | 47.0 | |
| **8** | a1vcnhxeij | 2010-01-04 | 20100104004211 | 2010-07-29 | FEMALE | 50.0 | |
| **9** | 6uh8zyj2gn | 2010-01-04 | 20100104023758 | 2010-01-04 | -unknown- | 46.0 | |
| **10** | yuuqmid2rp | 2010-01-04 | 20100104194251 | 2010-01-06 | FEMALE | 36.0 | |
| **11** | om1ss59ys8 | 2010-01-05 | 20100105051812 | NaN | FEMALE | 47.0 | |
| **12** | k6np330cm1 | 2010-01-05 | 20100105060859 | 2010-01-18 | -unknown- | NaN | |
| **13** | dy3rgx56cu | 2010-01-05 | 20100105083259 | NaN | FEMALE | 37.0 | |
| **14** | ju3h98ch3w | 2010-01-07 | 20100107055820 | NaN | FEMALE | 36.0 | |
| **15** | v4d5rl22px | 2010-01-07 | 20100107204555 | 2010-01-08 | FEMALE | 33.0 | |
| **16** | 2dwbwkx056 | 2010-01-07 | 20100107215125 | NaN | -unknown- | NaN | |
| **17** | frhre329au | 2010-01-07 | 20100107224625 | 2010-01-09 | -unknown- | 31.0 | |
| **18** | cxlg85pg1r | 2010-01-08 | 20100108015641 | NaN | -unknown- | NaN | |
| **19** | gdka1q5ktd | 2010-01-10 | 20100110010817 | 2010-01-10 | FEMALE | 29.0 | |

## Things to tidy up

Looking at the data we have,here are the areas we will be cleaning up

1.'date_account_created' and 'timestamp_first_active' data types will be converted to datetime

2.'date_first_booking' has a relatively low entries(a lot of missing values), may be dropped out.

3.'gender' has an inconsistent casing.

4.'age' has missing values. It will be considered for filling either with the mean or the median

5.'age' data type will be converted from float to integer

6.The 'age' column has some outliers

7.There are inconsistent casing in the 'first_device_type' and 'first_browser' column

In [4]:
```
# date_account_created converted from object to datetime
airbnbdata['date_account_created']= pd.to_datetime(airbnbdata['date_account_created'])
```

In [5]:
```
airbnbdata.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 213451 entries, 0 to 213450
Data columns (total 16 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   id                       213451 non-null  object
 1   date_account_created     213451 non-null  datetime64[ns]
 2   timestamp_first_active   213451 non-null  int64
 3   date_first_booking       88908 non-null   object
 4   gender                   213451 non-null  object
 5   age                      125461 non-null  float64
 6   signup_method            213451 non-null  object
 7   signup_flow              213451 non-null  int64
 8   language                 213451 non-null  object
 9   affiliate_channel        213451 non-null  object
 10  affiliate_provider       213451 non-null  object
 11  first_affiliate_tracked  207386 non-null  object
 12  signup_app               213451 non-null  object
 13  first_device_type        213451 non-null  object
 14  first_browser            213451 non-null  object
 15  country_destination      213451 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(2), object(12)
memory usage: 26.1+ MB
```

In [6]:
```
# timestamp_first_active converted from integer to datetime
airbnbdata['timestamp_first_active']= pd.to_datetime(airbnbdata['timestamp_first_active

airbnbdata.tail()
```

Out[6]:

| | id | date_account_created | timestamp_first_active | date_first_booking | gender | age |
|---|---|---|---|---|---|---|
| **213446** | zxodksqpep | 2014-06-30 | 2014-06-30 23:56:36 | NaN | MALE | 32.0 |
| **213447** | mhewnxesx9 | 2014-06-30 | 2014-06-30 23:57:19 | NaN | -unknown- | NaN |
| **213448** | 6o3arsjbb4 | 2014-06-30 | 2014-06-30 23:57:54 | NaN | -unknown- | 32.0 |
| **213449** | jh95kwisub | 2014-06-30 | 2014-06-30 23:58:22 | NaN | -unknown- | NaN |
| **213450** | nw9fwlyb5f | 2014-06-30 | 2014-06-30 23:58:24 | NaN | -unknown- | NaN |

In [7]:
```
#dropping the 'date_first_booking' column. It has a very high rate of missing values co
airbnbnew=airbnbdata.drop('date_first_booking', axis=1)
```

In [8]:
```
#changing the casing of the entries in this column to lowercase
airbnbnew['gender']=airbnbnew['gender'].str.lower()
```

In [9]:
```
#stripping off the hyphen symbol in the word '-unknown-'
airbnbnew['gender']=airbnbnew['gender'].str.strip('-')
airbnbnew
```

Out[9]:

| | id | date_account_created | timestamp_first_active | gender | age | signup_method | sign |
|---|---|---|---|---|---|---|---|
| **0** | gxn3p5htnn | 2010-06-28 | 2009-03-19 04:32:55 | unknown | NaN | facebook | |
| **1** | 820tgsjxq7 | 2011-05-25 | 2009-05-23 17:48:09 | male | 38.0 | facebook | |
| **2** | 4ft3gnwmtx | 2010-09-28 | 2009-06-09 23:12:47 | female | 56.0 | basic | |
| **3** | bjjt8pjhuk | 2011-12-05 | 2009-10-31 06:01:29 | female | 42.0 | facebook | |
| **4** | 87mebub9p4 | 2010-09-14 | 2009-12-08 06:11:05 | unknown | 41.0 | basic | |
| **...** | ... | ... | ... | ... | ... | ... | |
| **213446** | zxodksqpep | 2014-06-30 | 2014-06-30 23:56:36 | male | 32.0 | basic | |
| **213447** | mhewnxesx9 | 2014-06-30 | 2014-06-30 23:57:19 | unknown | NaN | basic | |
| **213448** | 6o3arsjbb4 | 2014-06-30 | 2014-06-30 23:57:54 | unknown | 32.0 | basic | |
| **213449** | jh95kwisub | 2014-06-30 | 2014-06-30 23:58:22 | unknown | NaN | basic | |
| **213450** | nw9fwlyb5f | 2014-06-30 | 2014-06-30 23:58:24 | unknown | NaN | basic | |

213451 rows × 15 columns

In [10]:
```
#checking the statistical values of the 'age' column
airbnbnew.describe()
```

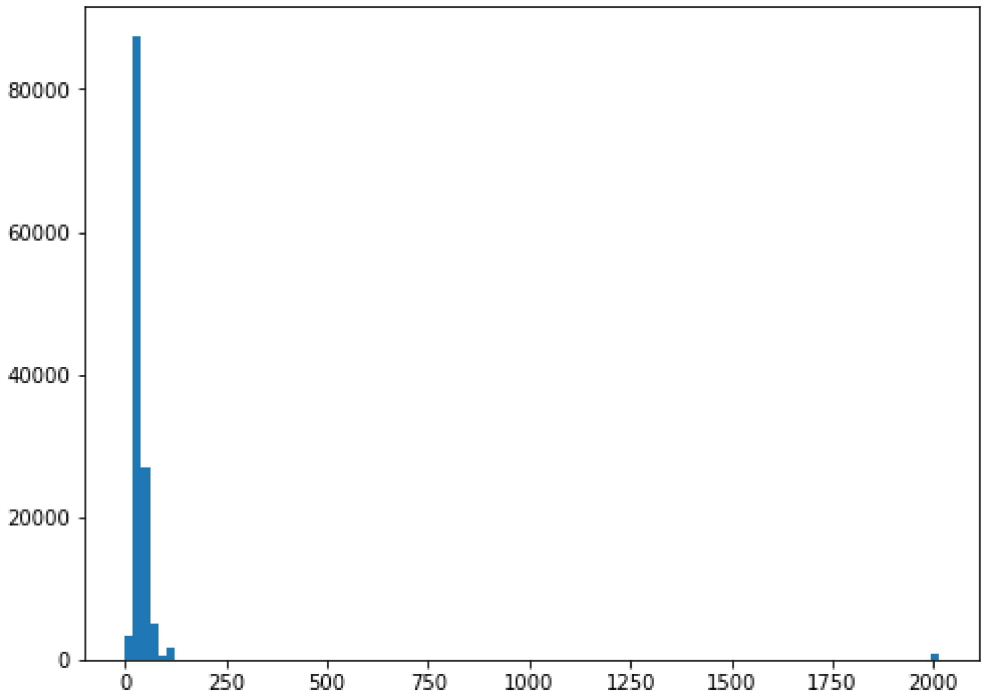
Out[10]:

| | age | signup_flow |
|---|---|---|
| **count** | 125461.000000 | 213451.000000 |
| **mean** | 49.668335 | 3.267387 |
| **std** | 155.666612 | 7.637707 |
| **min** | 1.000000 | 0.000000 |
| **25%** | 28.000000 | 0.000000 |
| **50%** | 34.000000 | 0.000000 |
| **75%** | 43.000000 | 0.000000 |
| **max** | 2014.000000 | 25.000000 |

In [11]:
```
fig = plt.figure()
```

```
plt.figure(figsize=(8,6))
plt.hist(airbnbdata['age'],bins=100)
plt.show()
```

```
<Figure size 432x288 with 0 Axes>
```




In [ ]:

In [12]:

```
#detected an unexpected and incorrect age in the 'age column'
#it looks like after the age 115, every other entry were years, so now we will calculat
#outage= airbnbnew[airbnbdata['age'] == 116]
#outage= airbnbnew[airbnbdata['age'] == 120]
maxage= airbnbnew[airbnbdata['age'] == 115]

maxage.head(30)
```

Out[12]:

| | id | date_account_created | timestamp_first_active | gender | age | signup_method | sig |
|---|---|---|---|---|---|---|---|
| **2228** | i0j7vqzk2m | 2010-10-29 | 2010-10-29 18:24:48 | male | 115.0 | facebook | |
| **9645** | gj79z2s4g4 | 2011-09-17 | 2011-09-17 13:14:29 | male | 115.0 | facebook | |
| **11530** | ai8c9x7w2l | 2011-10-22 | 2011-10-22 19:37:17 | male | 115.0 | facebook | |
| **15815** | 7cqnukdc9n | 2012-01-25 | 2012-01-25 04:06:51 | unknown | 115.0 | basic | |
| **36237** | 8u3tl9jtdl | 2012-08-13 | 2012-08-13 04:24:39 | male | 115.0 | facebook | |
| **39083** | fkczuwikaq | 2012-09-01 | 2012-09-01 16:48:22 | female | 115.0 | facebook | |
| **50539** | e8vpql9dyv | 2012-11-30 | 2012-11-30 19:31:54 | female | 115.0 | facebook | |
| **52115** | vecht936du | 2012-12-12 | 2012-12-12 22:00:14 | male | 115.0 | facebook | |

| | id | date_account_created | timestamp_first_active | gender | age | signup_method | sig |
|---|---|---|---|---|---|---|---|
| **52963** | 32sirobk34 | 2012-12-20 | 2012-12-20 17:52:22 | female | 115.0 | facebook | |
| **54803** | bo391whvc8 | 2013-01-06 | 2013-01-06 21:54:50 | male | 115.0 | facebook | |
| **59322** | 8n59s0i6gl | 2013-02-06 | 2013-02-06 20:04:42 | male | 115.0 | facebook | |
| **160420** | 8vjdu1l76i | 2014-03-07 | 2014-03-07 06:18:02 | female | 115.0 | basic | |
| **203271** | e1x7ej3lma | 2014-06-11 | 2014-06-11 22:37:50 | female | 115.0 | basic | |

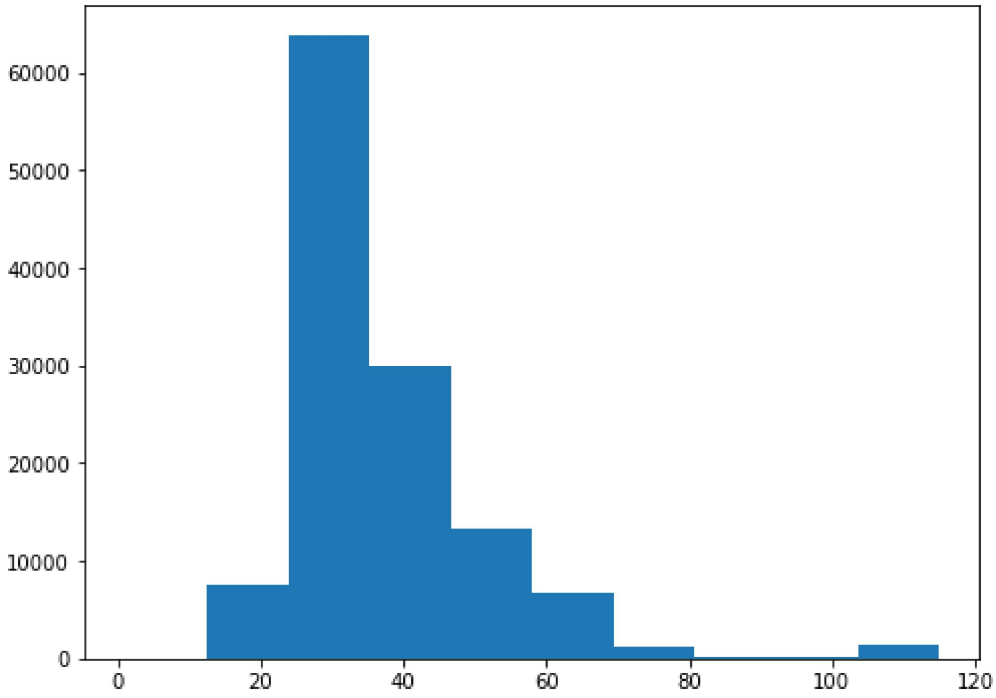
In [13]:
```
#let's drop those ages greater than 115 for now, when we work on them, we will add them
newairbnb= airbnbnew.loc[airbnbnew['age'] <=115]
```

In [14]:
```
oddage=airbnbnew.loc[airbnbnew['age'] >115]
#oddage['age']= 2014-oddage[['age']]
#oddage.head()
oddage.info()
#after converting them, it is clear and justifiable to declare them as wrong entries,
#therefore we'll be dropping it entirely and our new working dataframe will be 'newairb
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 781 entries, 388 to 211496
Data columns (total 15 columns):
 #   Column                   Non-Null Count  Dtype
---  ------                   --------------  -----
 0   id                       781 non-null    object
 1   date_account_created     781 non-null    datetime64[ns]
 2   timestamp_first_active   781 non-null    datetime64[ns]
 3   gender                   781 non-null    object
 4   age                      781 non-null    float64
 5   signup_method            781 non-null    object
 6   signup_flow              781 non-null    int64
 7   language                 781 non-null    object
 8   affiliate_channel        781 non-null    object
 9   affiliate_provider       781 non-null    object
 10  first_affiliate_tracked  747 non-null    object
 11  signup_app               781 non-null    object
 12  first_device_type        781 non-null    object
 13  first_browser            781 non-null    object
 14  country_destination      781 non-null    object
dtypes: datetime64[ns](2), float64(1), int64(1), object(11)
memory usage: 97.6+ KB
```

In [15]:
```
fig = plt.figure()
plt.figure(figsize=(8,6))
plt.hist(newairbnb['age'],bins=10)
plt.show()
```

```
<Figure size 432x288 with 0 Axes>
```




In [16]:
```
print(newairbnb['age'].max())
```

```
115.0
```

In [17]:
```
newairbnb.describe()
```

Out[17]:

| | age | signup_flow |
|---|---|---|
| **count** | 124680.000000 | 124680.000000 |
| **mean** | 37.411870 | 2.517717 |
| **std** | 13.952402 | 6.568004 |
| **min** | 1.000000 | 0.000000 |
| **25%** | 28.000000 | 0.000000 |
| **50%** | 34.000000 | 0.000000 |
| **75%** | 43.000000 | 0.000000 |
| **max** | 115.000000 | 25.000000 |

In [18]:
```
#replacing the missing values with the median of the entire age population after much c
newairbnb['age']=newairbnb['age'].fillna(newairbnb['age'].mean())
```

In [19]:
```
newairbnb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 124680 entries, 1 to 213448
Data columns (total 15 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
```

```
 0   id                       124680 non-null  object
 1   date_account_created     124680 non-null  datetime64[ns]
 2   timestamp_first_active   124680 non-null  datetime64[ns]
 3   gender                   124680 non-null  object
 4   age                      124680 non-null  float64
 5   signup_method            124680 non-null  object
 6   signup_flow              124680 non-null  int64
 7   language                 124680 non-null  object
 8   affiliate_channel        124680 non-null  object
 9   affiliate_provider       124680 non-null  object
 10  first_affiliate_tracked  122682 non-null  object
 11  signup_app               124680 non-null  object
 12  first_device_type        124680 non-null  object
 13  first_browser            124680 non-null  object
 14  country_destination      124680 non-null  object
dtypes: datetime64[ns](2), float64(1), int64(1), object(11)
memory usage: 15.2+ MB
```

In [20]:

```
newairbnb['age']=newairbnb['age'].astype('int64')
newairbnb.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 124680 entries, 1 to 213448
Data columns (total 15 columns):
 #   Column                   Non-Null Count   Dtype
---  ------                   --------------   -----
 0   id                       124680 non-null  object
 1   date_account_created     124680 non-null  datetime64[ns]
 2   timestamp_first_active   124680 non-null  datetime64[ns]
 3   gender                   124680 non-null  object
 4   age                      124680 non-null  int64
 5   signup_method            124680 non-null  object
 6   signup_flow              124680 non-null  int64
 7   language                 124680 non-null  object
 8   affiliate_channel        124680 non-null  object
 9   affiliate_provider       124680 non-null  object
 10  first_affiliate_tracked  122682 non-null  object
 11  signup_app               124680 non-null  object
 12  first_device_type        124680 non-null  object
 13  first_browser            124680 non-null  object
 14  country_destination      124680 non-null  object
dtypes: datetime64[ns](2), int64(2), object(11)
memory usage: 15.2+ MB
```

In [21]:

```
newairbnb.head(35)
```

Out[21]:

| | id | date_account_created | timestamp_first_active | gender | age | signup_method | signup_fl |
|---|---|---|---|---|---|---|---|
| **1** | 820tgsjxq7 | 2011-05-25 | 2009-05-23 17:48:09 | male | 38 | facebook | |
| **2** | 4ft3gnwmtx | 2010-09-28 | 2009-06-09 23:12:47 | female | 56 | basic | |
| **3** | bjjt8pjhuk | 2011-12-05 | 2009-10-31 06:01:29 | female | 42 | facebook | |
| **4** | 87mebub9p4 | 2010-09-14 | 2009-12-08 06:11:05 | unknown | 41 | basic | |
| **6** | lsw9q7uk0j | 2010-01-02 | 2010-01-02 01:25:58 | female | 46 | basic | |

| | id | date_account_created | timestamp_first_active | gender | age | signup_method | signup_fl |
|---|---|---|---|---|---|---|---|
| **7** | 0d01nltbrs | 2010-01-03 | 2010-01-03 19:19:05 | female | 47 | basic | |
| **8** | a1vcnhxeij | 2010-01-04 | 2010-01-04 00:42:11 | female | 50 | basic | |
| **9** | 6uh8zyj2gn | 2010-01-04 | 2010-01-04 02:37:58 | unknown | 46 | basic | |
| **10** | yuuqmid2rp | 2010-01-04 | 2010-01-04 19:42:51 | female | 36 | basic | |
| **11** | om1ss59ys8 | 2010-01-05 | 2010-01-05 05:18:12 | female | 47 | basic | |
| **13** | dy3rgx56cu | 2010-01-05 | 2010-01-05 08:32:59 | female | 37 | basic | |
| **14** | ju3h98ch3w | 2010-01-07 | 2010-01-07 05:58:20 | female | 36 | basic | |
| **15** | v4d5rl22px | 2010-01-07 | 2010-01-07 20:45:55 | female | 33 | basic | |
| **17** | frhre329au | 2010-01-07 | 2010-01-07 22:46:25 | unknown | 31 | basic | |
| **19** | gdka1q5ktd | 2010-01-10 | 2010-01-10 01:08:17 | female | 29 | basic | |
| **21** | qsibmuz9sx | 2010-01-10 | 2010-01-10 22:09:41 | male | 30 | basic | |
| **22** | 80f7dwscrn | 2010-01-11 | 2010-01-11 03:14:38 | unknown | 40 | basic | |
| **24** | 7i49vnuav6 | 2010-01-11 | 2010-01-11 23:08:08 | female | 40 | basic | |
| **25** | al8bcetz0g | 2010-01-12 | 2010-01-12 13:14:44 | female | 26 | basic | |
| **27** | hfrl5gle36 | 2010-01-12 | 2010-01-12 20:59:49 | female | 32 | basic | |
| **28** | tp6x3md0n4 | 2010-01-13 | 2010-01-13 04:46:50 | unknown | 35 | basic | |
| **29** | hql77nu2lk | 2010-01-13 | 2010-01-13 06:43:33 | unknown | 37 | basic | |
| **30** | cheova4spt | 2010-01-14 | 2010-01-14 02:52:57 | unknown | 42 | basic | |
| **31** | pggg2sj27u | 2010-01-14 | 2010-01-14 06:07:57 | male | 31 | basic | |
| **32** | mo2v5h3nti | 2010-01-14 | 2010-01-14 08:49:26 | unknown | 31 | basic | |
| **33** | qthj88nnc7 | 2010-01-14 | 2010-01-14 17:52:09 | male | 29 | basic | |
| **34** | x9gnxoun57 | 2010-01-15 | 2010-01-15 05:50:53 | male | 59 | basic | |
| **35** | ugy4obax11 | 2010-01-15 | 2010-01-15 08:27:11 | unknown | 49 | basic | |
| **36** | 7my0vrljxc | 2010-01-15 | 2010-01-15 22:19:29 | female | 31 | basic | |
| **37** | ul0b29nhr6 | 2010-01-15 | 2010-01-15 23:16:51 | male | 30 | basic | |
| **39** | myunxar49t | 2010-01-16 | 2010-01-16 15:49:42 | female | 35 | basic | |
| **40** | jpne62dhz8 | 2010-01-16 | 2010-01-16 21:59:43 | male | 26 | basic | |
| **41** | k15j7mbny0 | 2010-01-19 | 2010-01-19 01:36:16 | female | 30 | basic | |
| **42** | 2981o5kr7b | 2010-01-21 | 2010-01-21 07:09:39 | female | 29 | basic | |
| **45** | g1q6caq452 | 2010-01-23 | 2010-01-23 20:15:40 | female | 44 | basic | |

In [22]:
```
#taking out unwanted symbols
newairbnb['first_device_type'].unique()
```

Out[22]:
```
array(['Mac Desktop', 'Windows Desktop', 'iPhone', 'Other/Unknown',
       'Desktop (Other)', 'Android Tablet', 'iPad', 'Android Phone',
       'SmartPhone (Other)'], dtype=object)
```

In [23]:
```
newairbnb['first_device_type']=newairbnb['first_device_type'].str.replace('Other/','')
```

In [24]:
```
newairbnb['first_device_type'].unique()
```

Out[24]:
```
array(['Mac Desktop', 'Windows Desktop', 'iPhone', 'Unknown',
       'Desktop (Other)', 'Android Tablet', 'iPad', 'Android Phone',
       'SmartPhone (Other)'], dtype=object)
```

In [25]:
```
#taking out unwanted symbols
newairbnb['first_browser']=newairbnb['first_browser'].str.strip('-')
newairbnb['first_browser'].unique()
```

Out[25]:
```
array(['Chrome', 'IE', 'Firefox', 'Safari', 'unknown', 'Mobile Safari',
       'Chrome Mobile', 'RockMelt', 'Chromium', 'Android Browser',
       'AOL Explorer', 'Mobile Firefox', 'Opera', 'TenFourFox',
       'Apple Mail', 'Silk', 'Camino', 'IE Mobile', 'BlackBerry Browser',
       'SeaMonkey', 'Sogou Explorer', 'IceWeasel', 'SiteKiosk',
       'Opera Mini', 'Maxthon', 'Kindle Browser', 'CoolNovo',
       'wOSBrowser', 'Iron', 'Mozilla', 'PS Vita browser', 'NetNewsWire',
       'Pale Moon', 'Avant Browser', 'Opera Mobile', 'Yandex.Browser',
       'CometBird', 'TheWorld Browser', 'SlimBrowser', 'Comodo Dragon',
       'Stainless'], dtype=object)
```

In [ ]: