airbnb

Airbnb New User Bookings

Business School of Stevens Institute of Technology
Ke Cao

Introduction

- Instead of waking to overlooked "Do not disturb“ sign, Airbnb travelers find themselves rising with the birds in a whimsical treehouse, having their morning coffee on the deck of a houseboat, or cooking a shared regional breakfast with their hosts.

- New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries. By accurately predicting where a new user will book their first travel experience, Airbnb can share more personalized content with their community, decrease the average time to first booking, and better forecast demand.

Data and Objective

<DATA SOURCE>---

- https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data

<OBJECTIVE>---

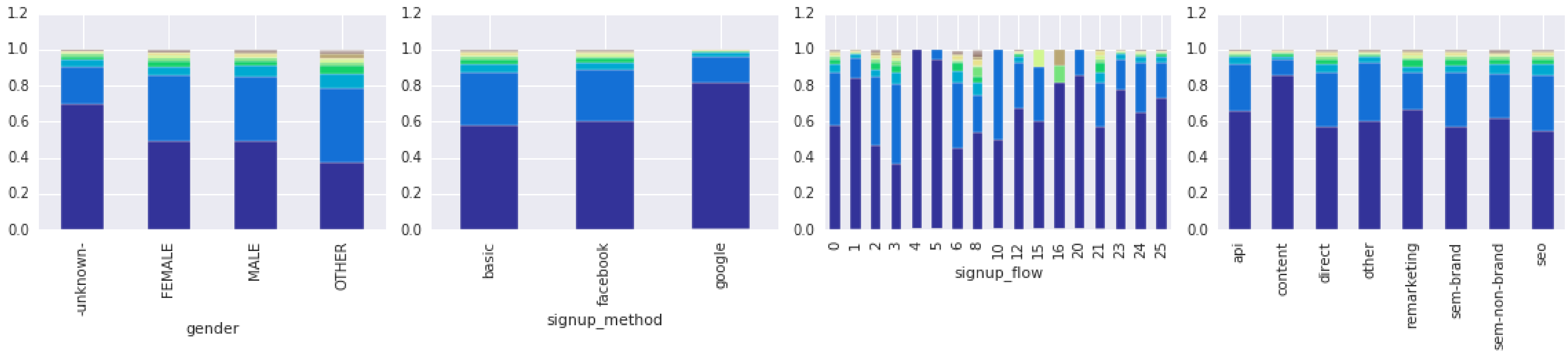
- To predict which country a new user's first booking destination will likely be

Data and Objective

	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
0	gxn3p5htnn	2010-06-28	20090319043255	NaN	-unknown-	NaN	facebook	0	en	direct	direct	untracked	Web	Mac Desktop	Chrome	NDF
1	820tgsjxq7	2011-05-25	20090523174809	NaN	MALE	38.0	facebook	0	en	seo	google	untracked	Web	Mac Desktop	Chrome	NDF
2	4ft3gnwmtx	2010-09-28	20090609231247	2010-08-02	FEMALE	56.0	basic	3	en	direct	direct	untracked	Web	Windows Desktop	IE	US
3	bjjt8pjhuk	2011-12-05	20091031060129	2012-09-08	FEMALE	42.0	facebook	0	en	direct	direct	untracked	Web	Mac Desktop	Firefox	other
4	87mebub9p4	2010-09-14	20091208061105	2010-02-18	-unknown-	41.0	basic	0	en	direct	direct	untracked	Web	Mac Desktop	Chrome	US

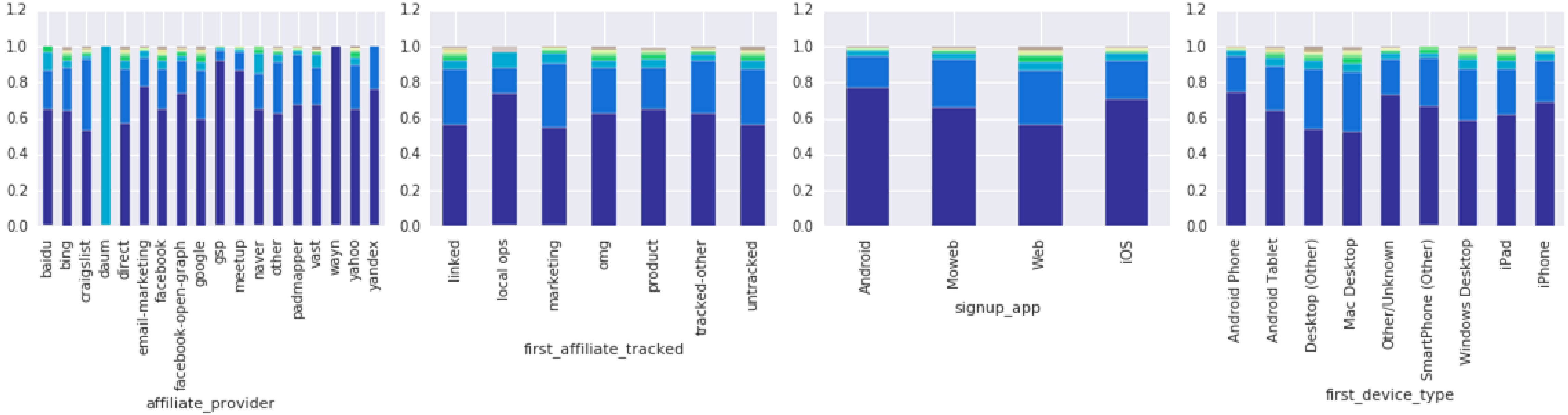
- There are **213451** observations in the train dataset.
- Features of data includes: **Gender, Age, Language, Country_destination, Signup_app, etc**.
- There are **12** possible outcomes of the destination country: **'US', 'FR', 'CA', 'GB', 'ES', 'IT', 'PT', 'NL','DE', 'AU', 'NDF'** (no destination found, means there wasn't a booking), and **'other'**.

A Summary of Characters in Dataset



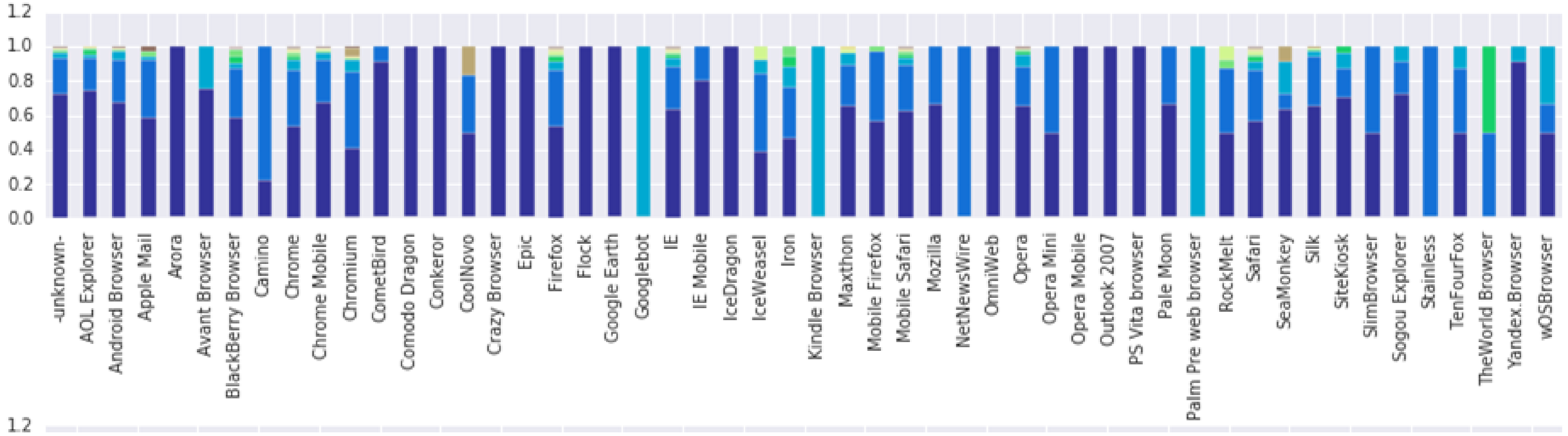
- Starting with **gender**, it appears users with 'unknown' gender book less frequently than those with a known one while users with gender 'other' book more frequently
- Users with the 'google' **signup_method** book less frequently than 'basic' or 'facebook'
- Users with **signup_flow---** ‘3’ book more frequently than any other category while several have nearly 100% 'NDF'
- Users with **affiliate_channel** 'content' book less frequently than other categories

A Summary of Characters in Dataset



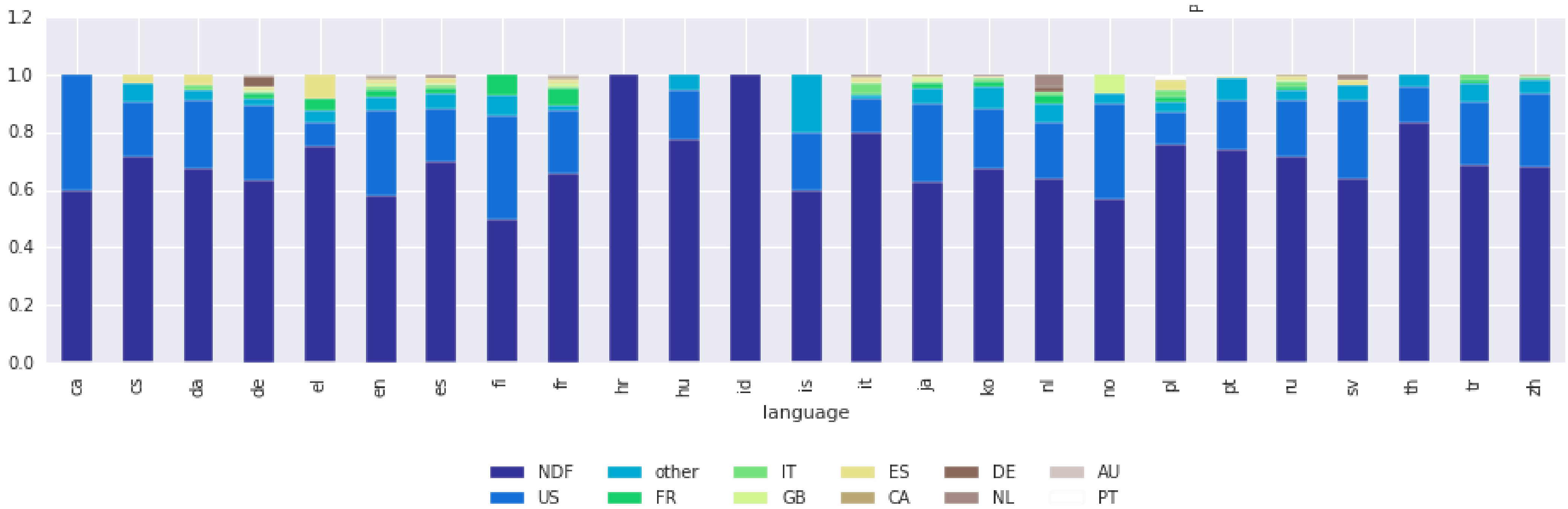
- Users with **affiliate_provider**---'**craigslist**', **direct**', and '**google**' book more frequently than other categories.
- Users with **first_affiliate_tracked---** '**local ops**' book less frequently than other categories.
- Users with **signup_app---** '**Web**' booked the most frequently, while those with 'Android' booked the least.
- Users with **first_device_type---** '**Mac_Desktop**' booked the most frequently, while those with 'Android Phone' booked the least.

A Summary of Characters in Dataset



- The chart on **first_browser** highlights the large number used above all else; it is difficult to achieve any meaningful insights beyond that some obscure browsers that are not likely widely used have very high or very low booking frequencies.

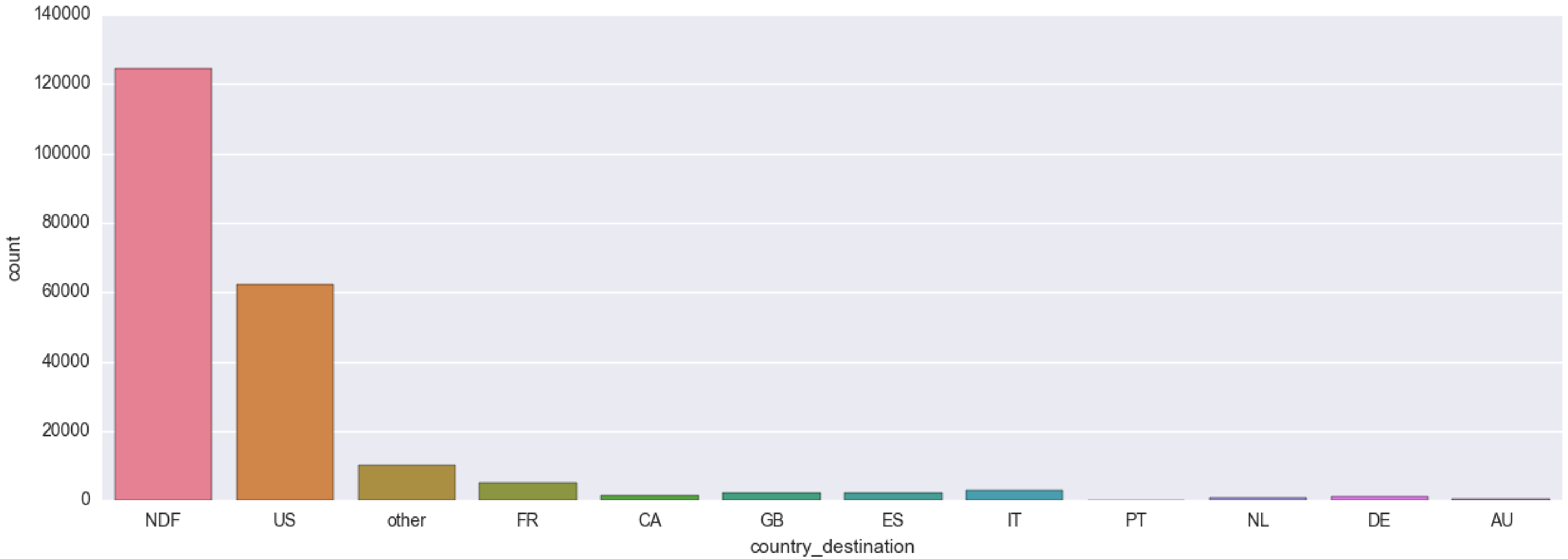
A Summary of Characters in Dataset



en	206314
zh	1632
fr	1172
es	915
ko	747
de	732
it	514
ru	389
pt	240
ja	225
sv	122
nl	97
tr	64
da	58
pl	54
cs	32
no	30
th	24
el	24
id	22
hu	18
fi	14
is	5
ca	5
hr	2

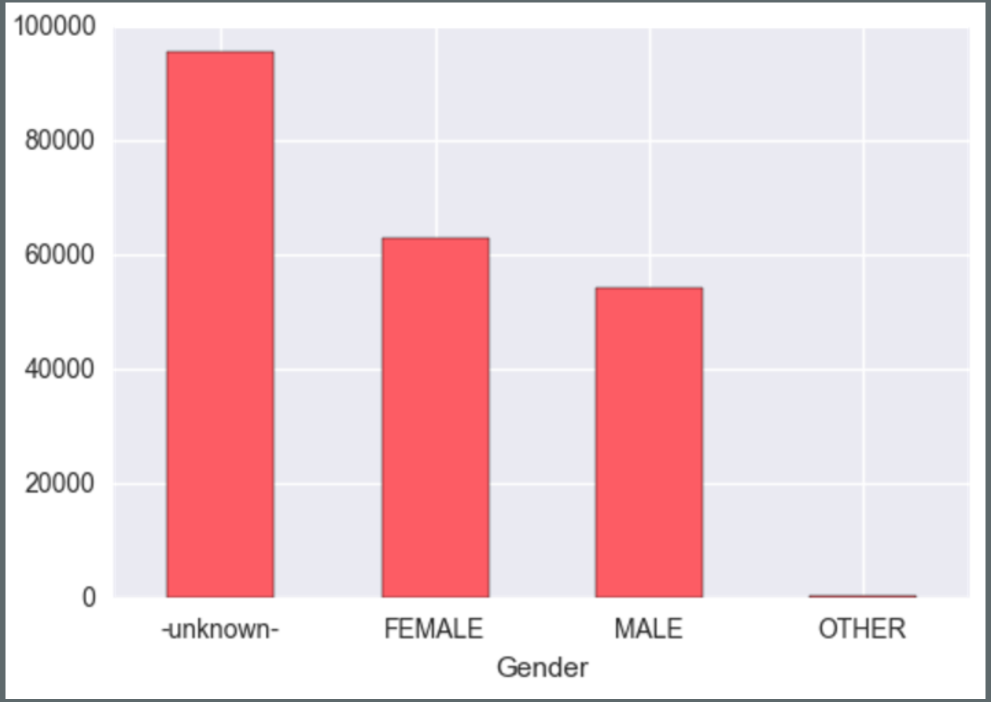
- The chart on **language** shows that most of the users were taking English as the main viewing language.

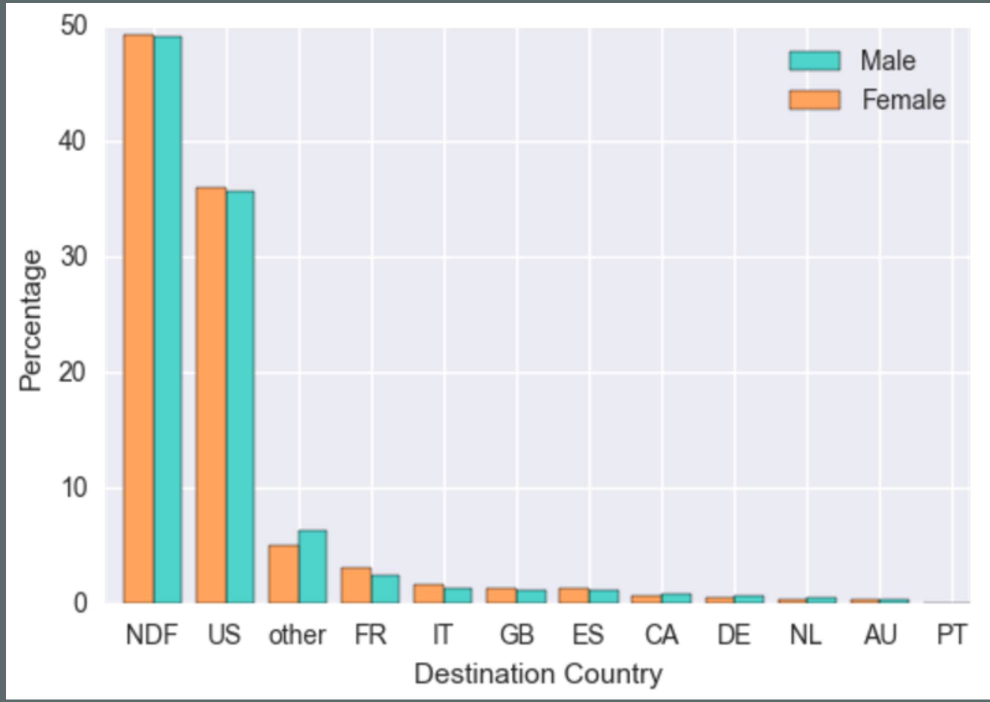
Value Count of Country_destination



Most users choose to travel to the U.S. , excluding those who registered but made no booking.

Consumer Behavior by Gender

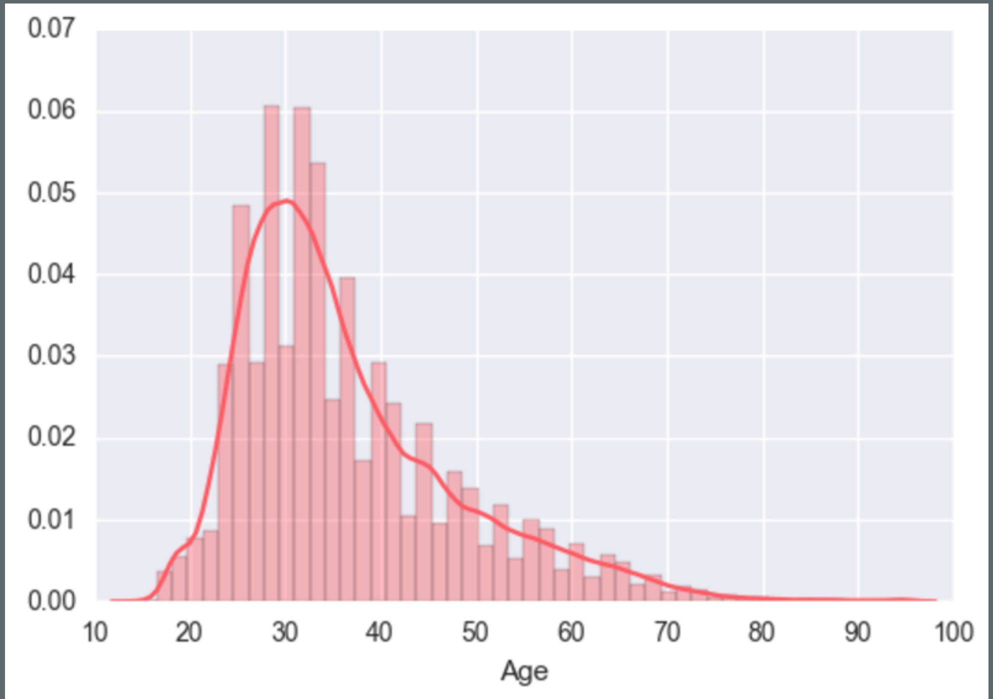


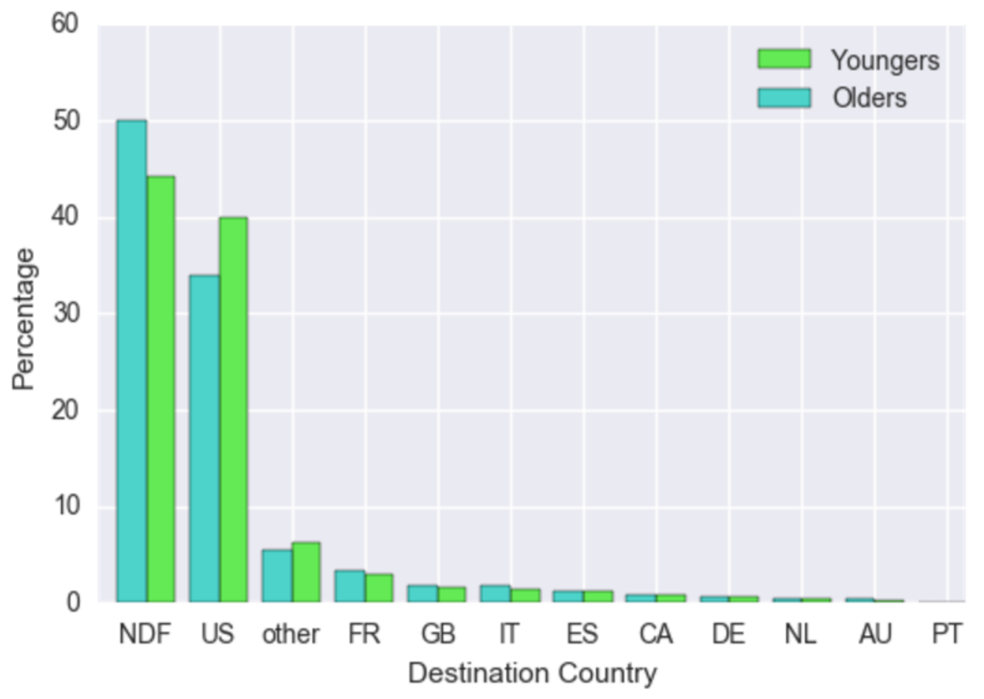


As we've seen before, at this plot we can see the amount of missing data in perspective. Also, notice that there is a slight difference between user gender.

Next thing is if there is any gender preferences.
There are no big differences between the 2 main genders, so it's not really useful except to know the relative destination frequency of the countries.

Consumer Behavior by Age





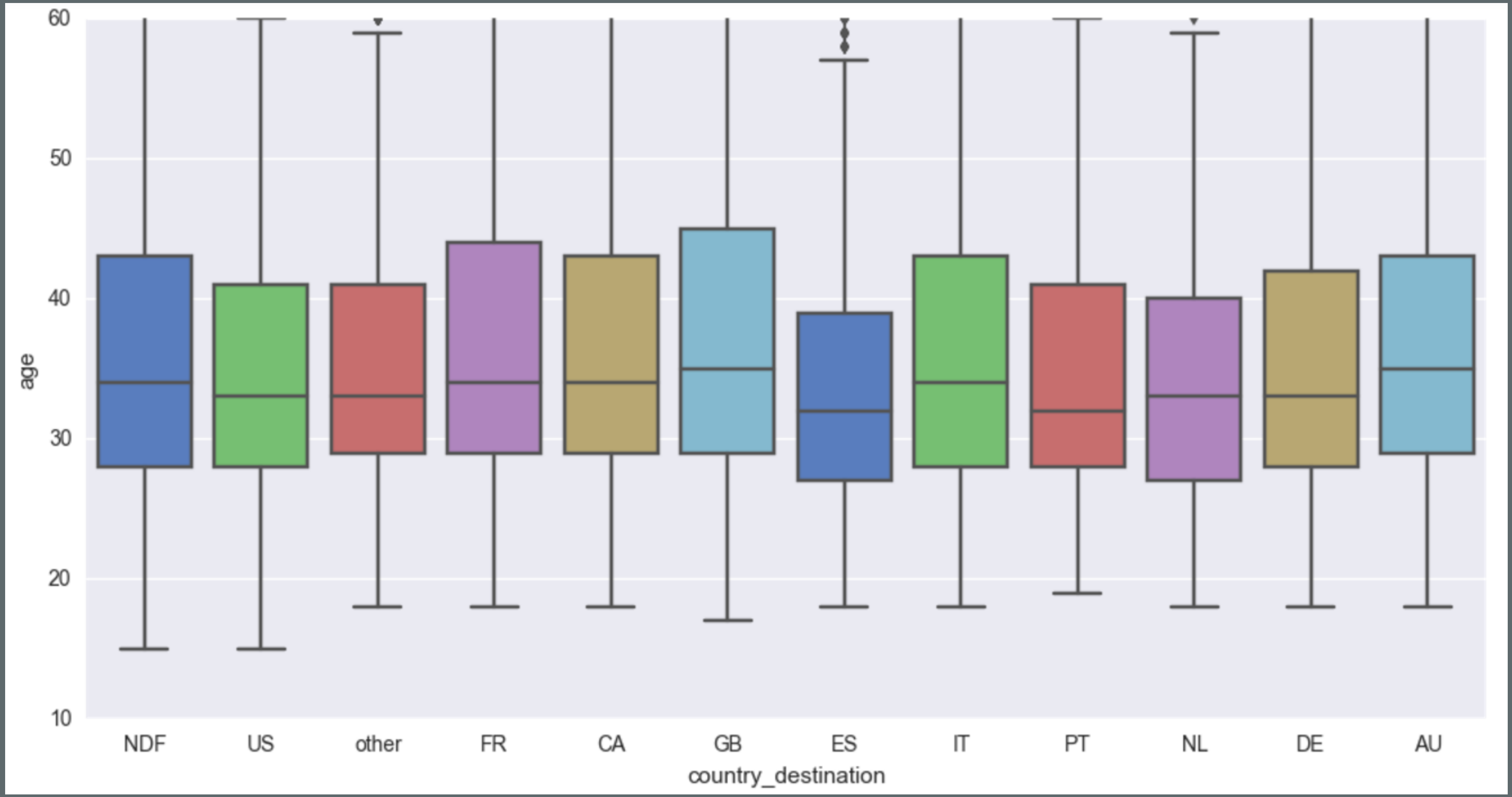
Digging into the age, generate a graph of frequency for age.

The common age to travel is between 25 and 40.

Then cut age values into groups, in this case we take 45 to vary youngers and olders.

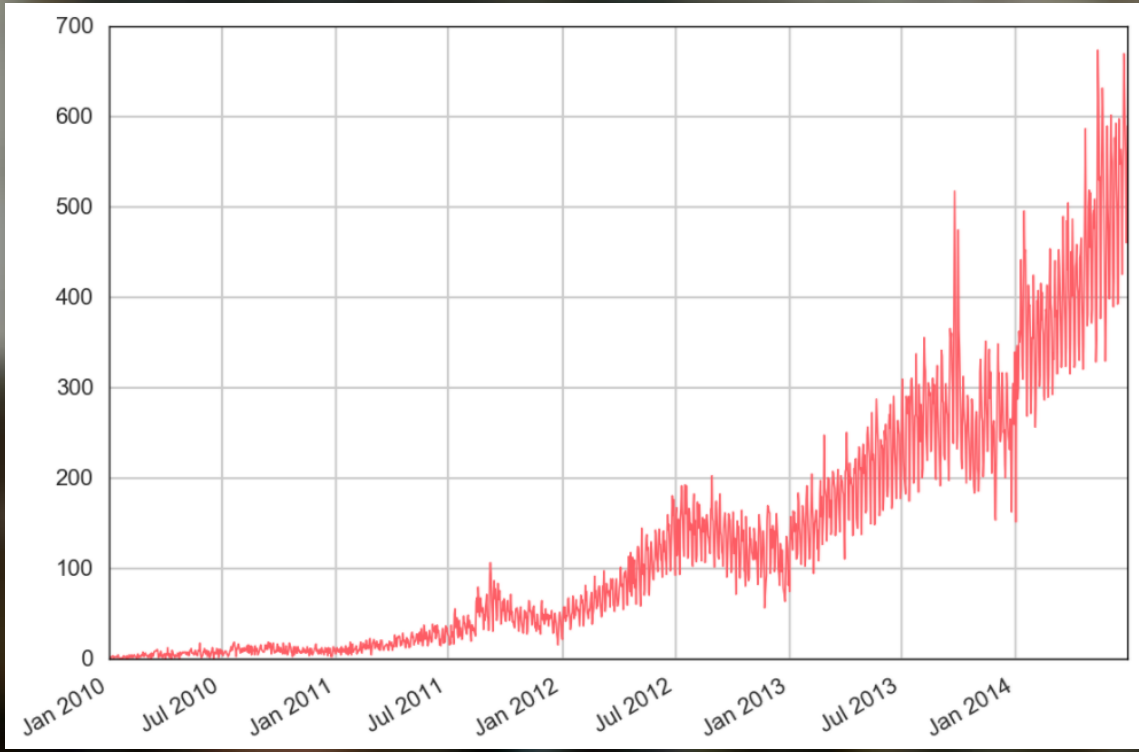
Young people tends to stay in the US, while the older people choose to travel outside the country (relatively).

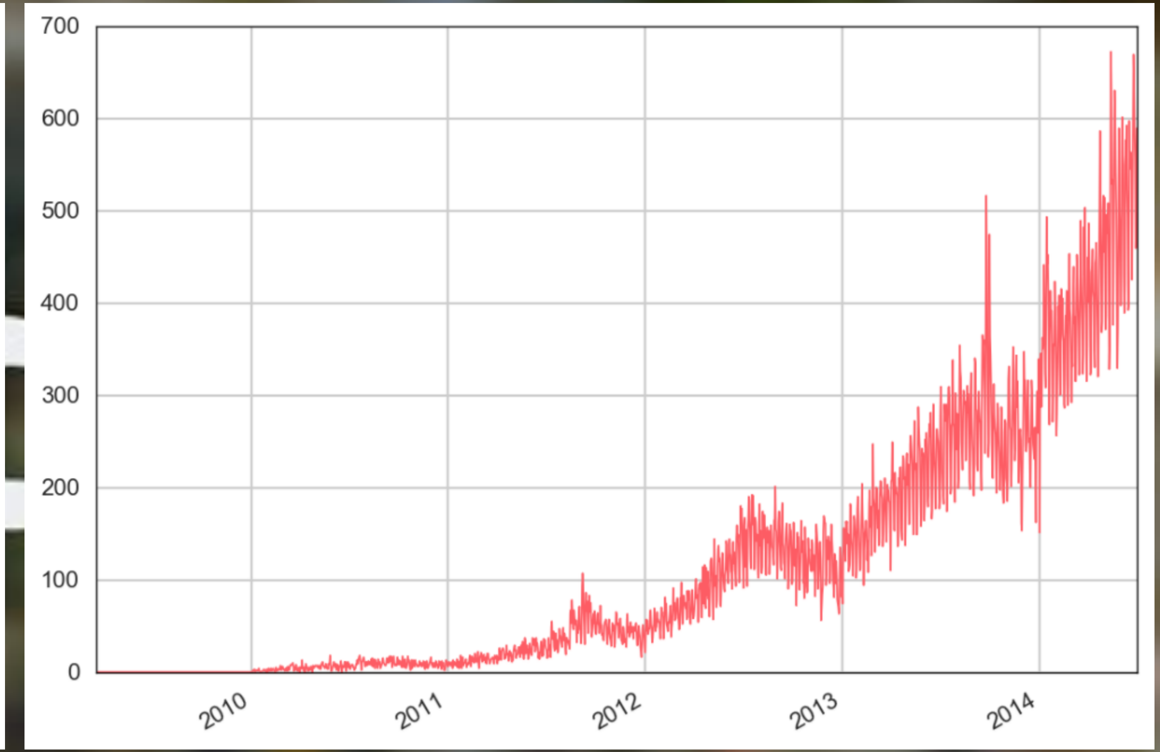
Consumer Behavior by Age



Users who book trips to Spain and Portugal tend to be younger while those that book trips to Great Britain tend to be older.

Consumer Behavior by Time



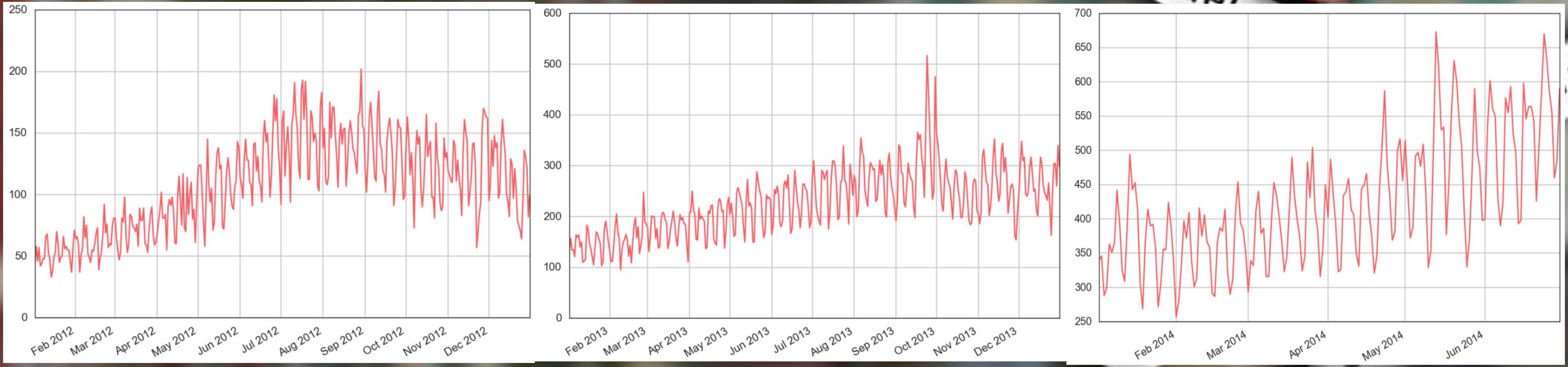


Plot the ***Number of Accounts Created by Time***.

Airbnb has grown fast over the 3 years(2012-2014).

Plot the ***Number of Users First Active by Time***.
We can see it's almost the same as ***"date_account_created"***, and also, notice the small peaks.

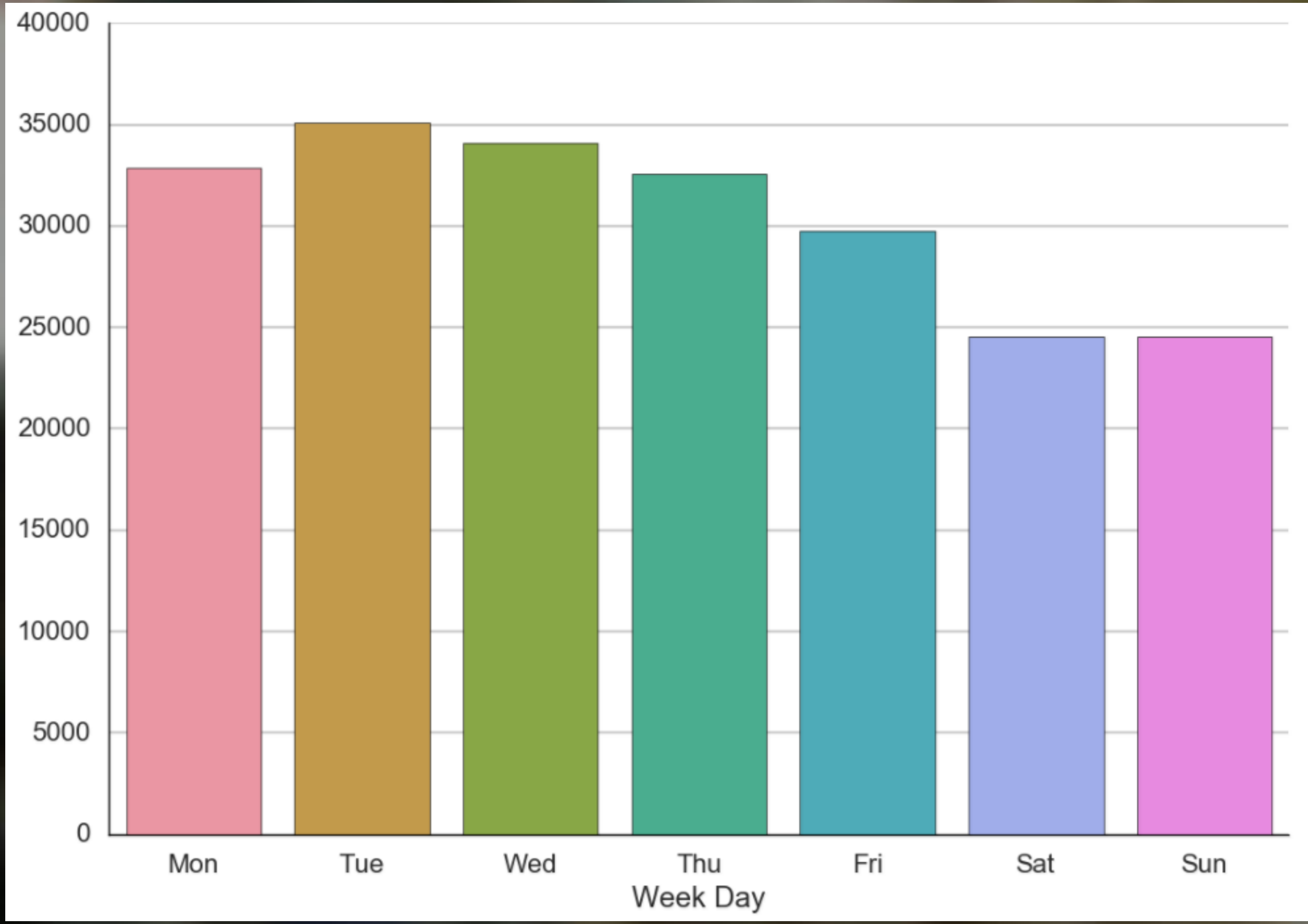
Consumer Behavior by Time



The most active months are---
July, August, and September.

There are some peaks at the same distance.

Consumer Behavior by Time



The minimum of accounts created lies on weekends(when people use less Internet), and day of accounts created usually hits a maximum on Tuesdays.

Destinations Clustering by Age

Airbnb Destination Clusters by Age

The FASTCLUS Procedure
Replace=FULL Radius=0 Maxclusters=5 Maxiter=1

Initial Seeds	
Cluster	age
1	15.0000000
2	57.0000000
3	36.0000000
4	78.0000000
5	100.0000000

```
data airbnb_clust;
set Airbnb;
if age>100 or age<10 then delete;
run;

proc fastclus data=airbnb_clust
maxclusters=5
out=destination_clusters;
var age;
id country_destination;
run;
```

Destinations Clustering by Age

Cluster Summary						
Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	31958	2.6150	7.8983		3	9.7735
2	21571	5.3050	9.6910		4	16.8503
3	65273	4.3139	10.1720		1	9.7735
4	3932	4.5116	11.0893		2	16.8503
5	325	4.8312	10.2778		4	23.4959

Cluster Means	
Cluster	age
1	25.07240753
2	52.02911316
3	34.84595468
4	68.87945066
5	92.37538462

Statistics for Variables				
Variable	Total STD	Within STD	R-Square	RSQ/(1-RSQ)
age	11.69061	4.15835	0.873481	6.903981
OVER-ALL	11.69061	4.15835	0.873481	6.903981

Pseudo F Statistic =	212390.6

Approximate Expected Over-All R-Squared =	0.96000

Cubic Clustering Criterion =	-299.621

Cluster Standard Deviations	
Cluster	age
1	2.614971835
2	5.304985439
3	4.313902785
4	4.511577002
5	4.831162692

Feature Engineering

```
In [92]: session = pd.read_csv('C://Users//amyhu//Google Drive//672//672finalprojext//sessions.csv//sessio

In [93]: session = session.rename(columns = {'user_id':'id'})

In [94]: session_tr = session[session.id.isin(id_train)]

In [95]: session_tr.head()
```

Out[95]:

	id	action	action_type	action_detail	device_type	secs_elapsed
0	d1mm9tcy42	lookup	NaN	NaN	Windows Desktop	319.0
1	d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	67753.0
2	d1mm9tcy42	lookup	NaN	NaN	Windows Desktop	301.0
3	d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	22141.0
4	d1mm9tcy42	lookup	NaN	NaN	Windows Desktop	435.0

```
In [96]: times = session_tr.id.value_counts().to_frame()

In [97]: type(times)

Out[97]: pandas.core.frame.DataFrame

In [98]: times = times.reset_index()

In [99]: times = times.rename(columns = {'id':'time','index':'id'})
         times.head()
```

- **Encode categorical features**: signup_method, signup_app, etc.
- **Split the time features**: signup_time, account created time, etc.
- **Generate customers behavior features**: number of visit times, average seconds spending on the website

Modeling

1. KNN

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=1, n_neighbors=5, p=2,
           weights='uniform')
```

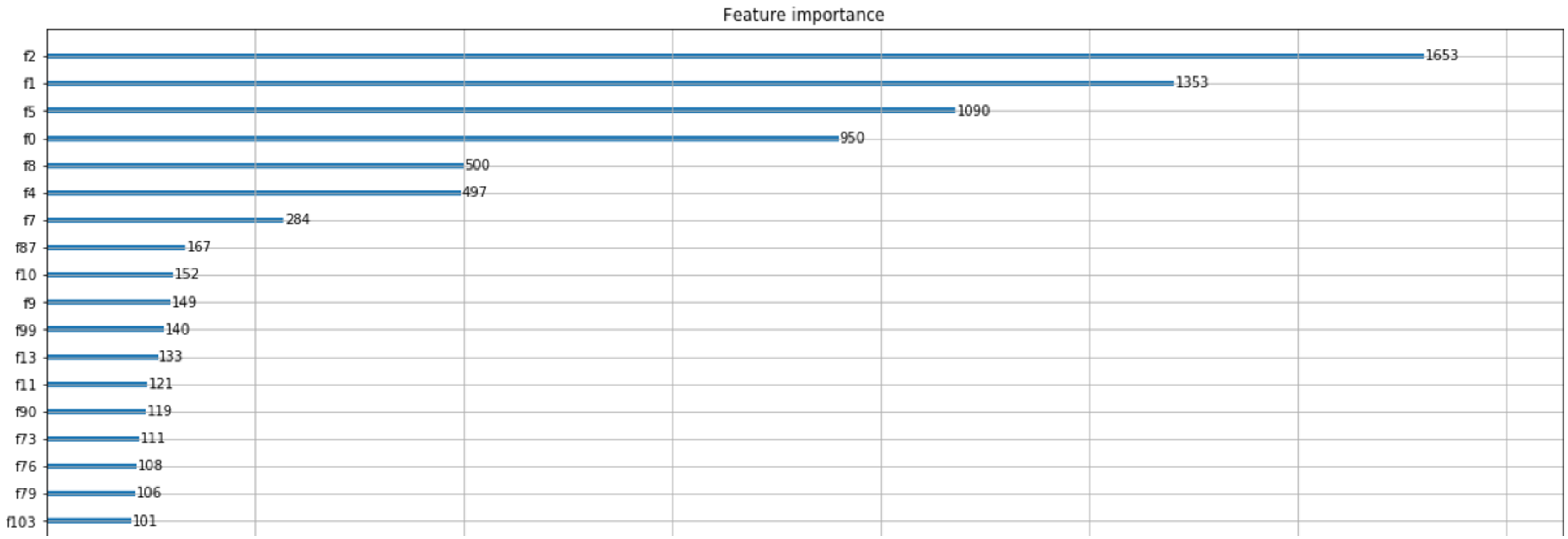
1. XGBoost

```
XGBClassifier(base_score=0.5, colsample_bylevel=1, colsample_bytree=0.5,
       gamma=0, learning_rate=0.3, max_delta_step=0, max_depth=6,
       min_child_weight=1, missing=None, n_estimators=25, nthread=-1,
       objective='multi:softprob', reg_alpha=0, reg_lambda=1,
       scale_pos_weight=1, seed=0, silent=True, subsample=0.5)
```

Logloss of XGB:0.99

Logloss of KNN: 5.18

Insight from XGBOOST

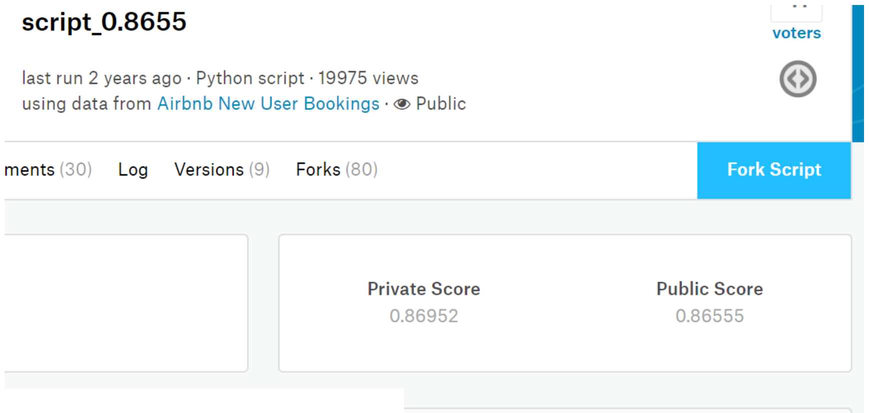


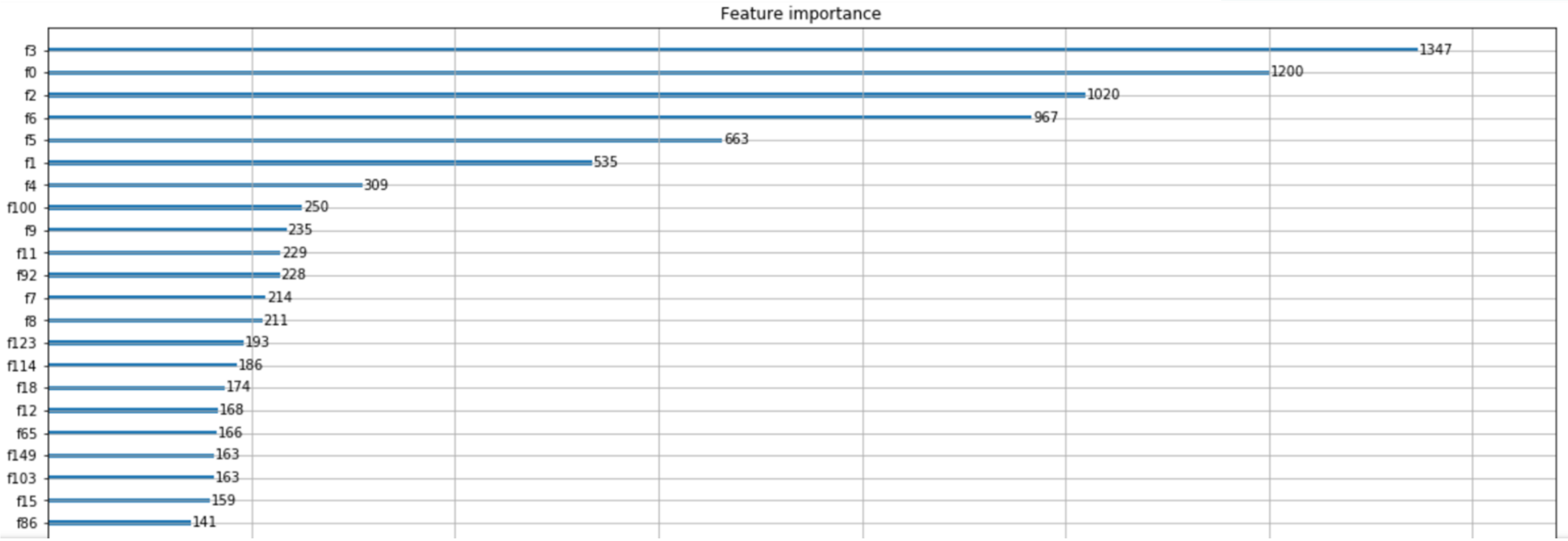
- **Features importance**

#1 average time spent on website   # 2 number of visit times

#3 the date of created account   #4 age   #5 the date of first time active

Bad Example: Loss Insight



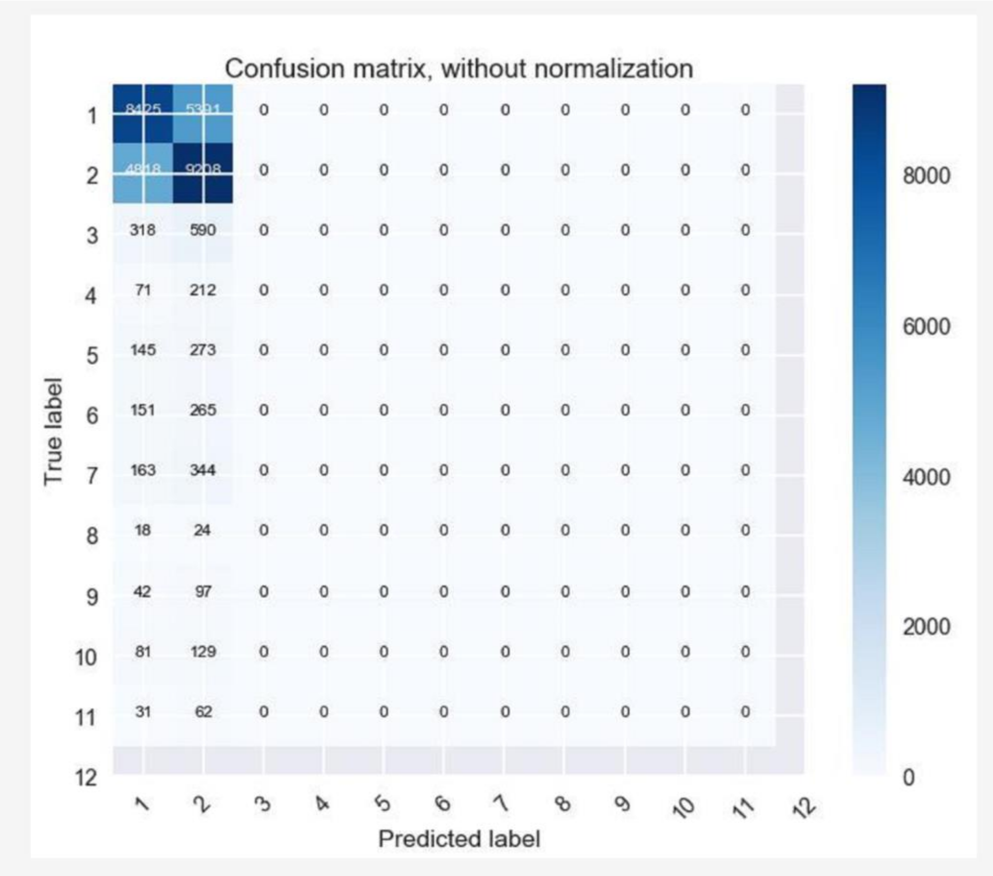


- Features importance

#1 the date of created account   # 2 age   #3 the month of created account

#4 the date of first time active   #5 the month of first time active

Confusion Matrix

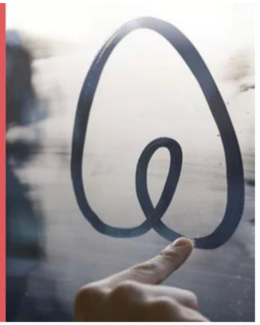


	country_destination
1	NDF
2	US
3	other
4	FR
5	CA
6	GB
7	ES
8	IT
9	PT
10	NL
11	DE
12	AU

In the confusion matrix, we can see that the trained support vector machine cannot classify the class from 3-12.

And the performance to classify the first two country destinations is better. That is verified that support vector machine is not sensitive to multi classification.

Suggestions of Marketing Campaign



- Tracking consumer behavior helps catch potential consumer. For example, Airbnb can invest more promotion budget to users who visit the website more than 1000 times.
- For homes, Airbnb may make recommendations for users in different ages, like recommend modern, convenient homes with complete entertainment facilities for youngers and quiet, comfortable homes with awesome views.
- For destinations, Airbnb may tag the alternatives with, for example, “Most Teenagers’ Choice”, “Friendly to The Aged”, “Girl’s Favorite”, “LGBT friendly”,etc.
- For attracting and retaining users, Airbnb can send emails to those not active users, to attract them visit our website, because the more visit the more likely they will book room. The emails can mostly be sent during Jul. to Sep., which is a period of “Travel Season”.


airbnb
THANKS!