Travel Predictions

AIRBNB USER BOOKING PREDICTIONS USING MACHINE LEARNING

Jeanette R. Henry

Regis University

Data Science Practicum II

Goal

Predict where new users will book a home with AirBnB

Using data about:

- User Profiles Information
- User Session Logs
- Statistics on age populations in countries

Tools

R programming

Main packages:

- xgboost: Machine learning model
- dply & data.table: Data manipulation and feature engineering

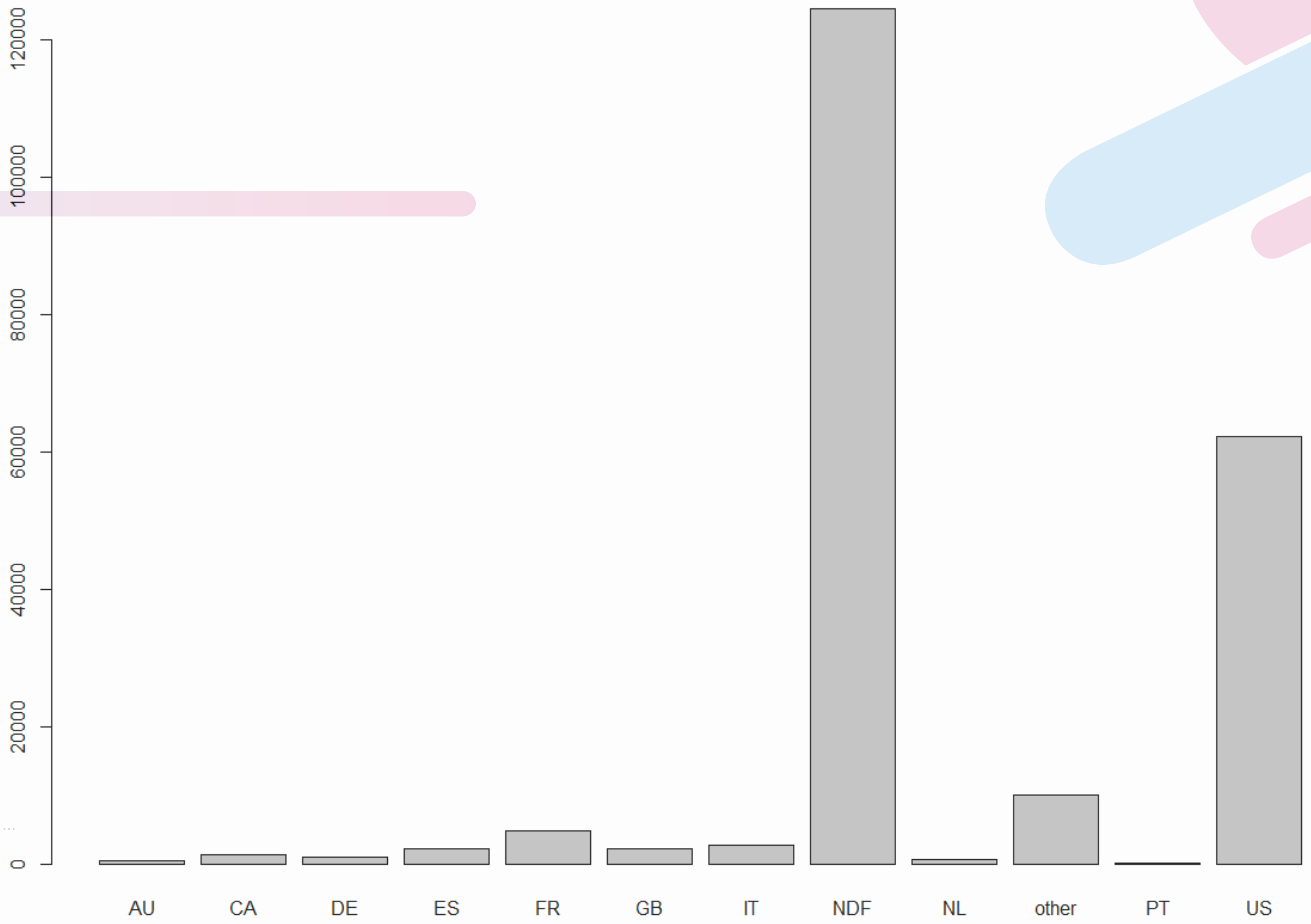
- Others:

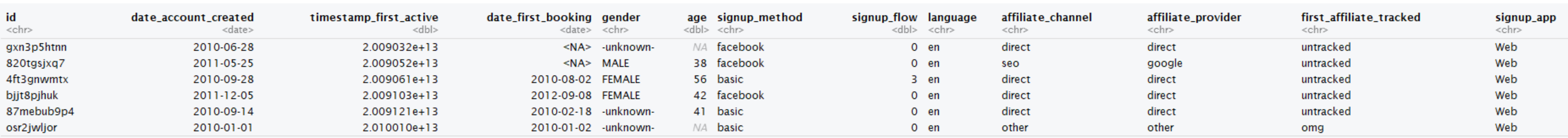
  readr,
  stringr,
  caret,
  car,
  DMwR,
  Hmisc,
  plyr,
  tidyr,
  DescTools,
  Matrix,
  glmnet,
  plotrix,
  Ckmeans.1d.dp,

User Data

User Data

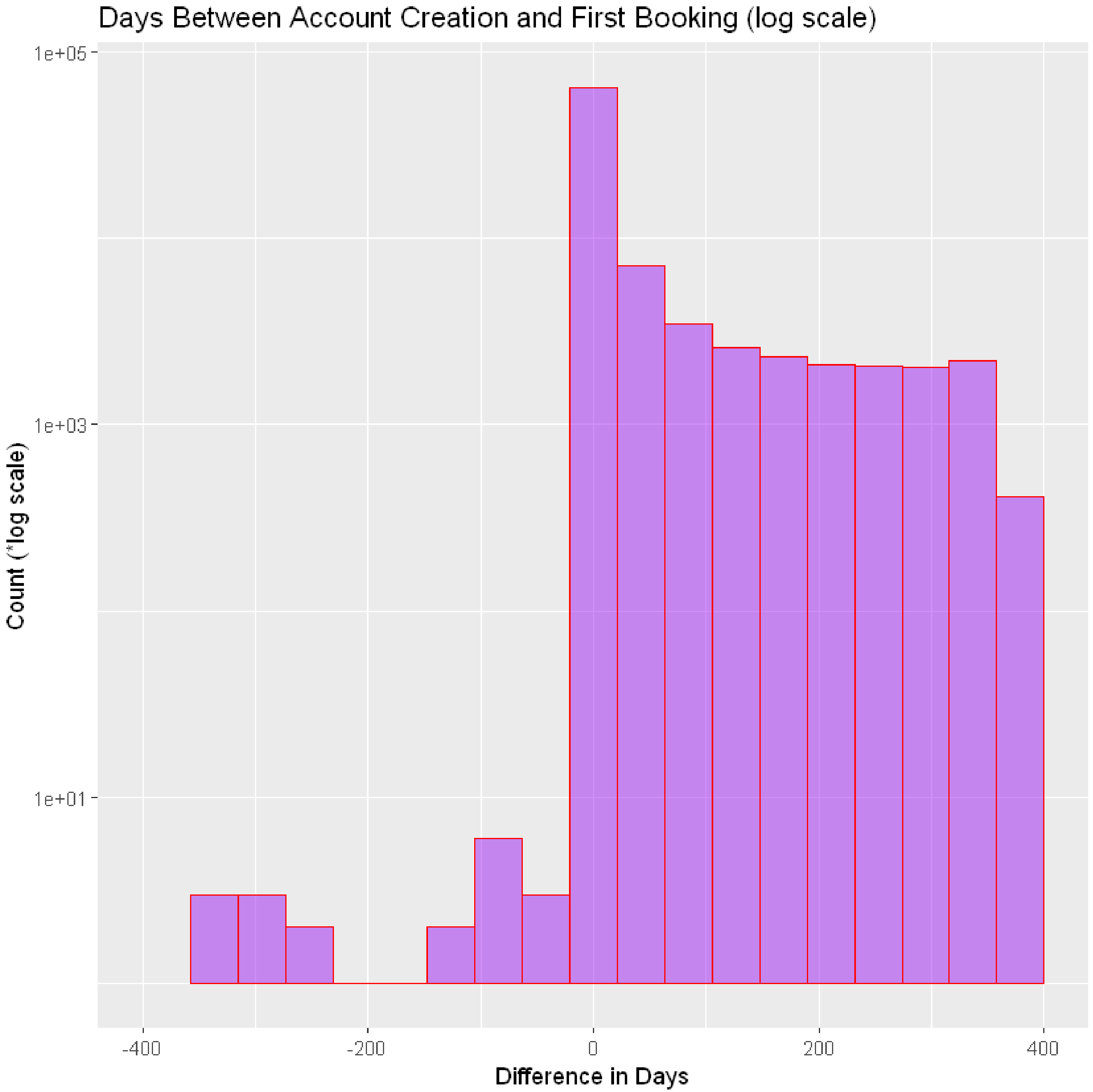
- 62096 obs. of 16 variables:
- $ id : chr "5uwns89zht" "jtl0dijy2j" "xx0ulgorjt" "6c6puo6ix0" ...
- $ date_account_created : Date, format: "2014-07-01" "2014-07-01" ...
- $ timestamp_first_active : num 2.01e+13 2.01e+13 2.01e+13 2.01e+13 2.01e+13 ...
- $ date_first_booking : logi NA NA NA NA NA NA ...
- $ gender : chr "FEMALE" "-unknown-" "-unknown-" "-unknown-" ...
- $ age : num 35 NA NA NA NA 28 48 NA NA NA ...
- $ signup_method : chr "facebook" "basic" "basic" "basic" ...
- $ signup_flow : num 0 0 0 0 0 0 25 0 0 0 ...
- $ language : chr "en" "en" "en" "en" ...
- $ affiliate_channel : chr "direct" "direct" "direct" "direct" ...
- $ affiliate_provider : chr "direct" "direct" "direct" "direct" ...
- $ first_affiliate_tracked: chr "untracked" "untracked" "linked" "linked" ...
- $ signup_app : chr "Moweb" "Moweb" "Web" "Web" ...
- $ first_device_type : chr "iPhone" "iPhone" "Windows Desktop" "Windows Desktop" ...
- $ first_browser : chr "Mobile Safari" "Mobile Safari" "Chrome" "IE" ...
- $ country_destination : logi NA NA NA NA NA NA ...



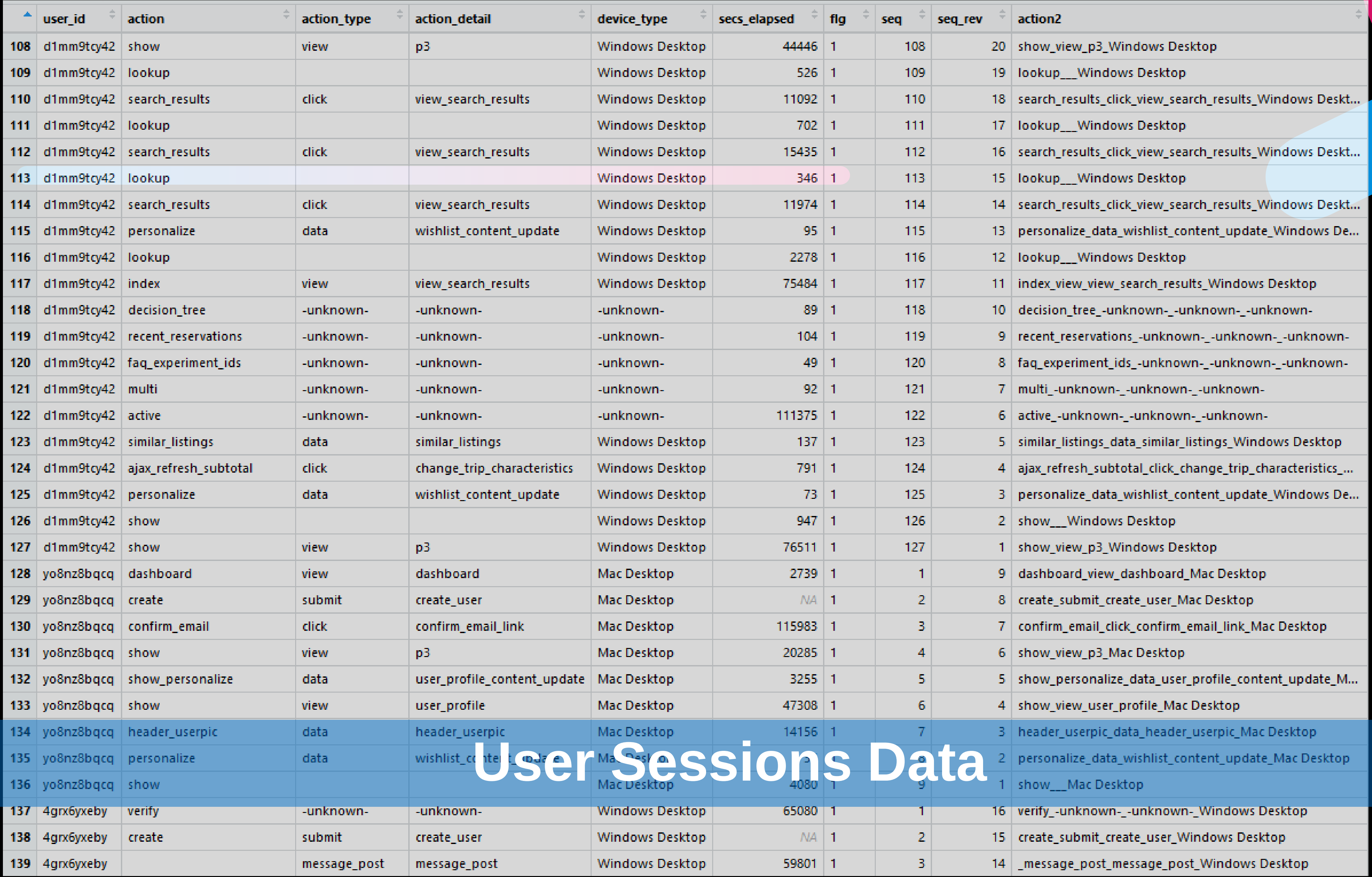


id	date_account_created	timestamp_first_active	date_first_booking	gender	age	signup_method	signup_flow	language	affiliate_channel	affiliate_provider	first_affiliate_tracked	signup_app
<chr>	<date>	<dbl>	<date>	<chr>	<dbl>	<chr>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>
gxn3p5htnn	2010-06-28	2.009032e+13	<NA>	-unknown-	NA	facebook	0	en	direct	direct	untracked	Web
820tgsjxq7	2011-05-25	2.009052e+13	<NA>	MALE	38	facebook	0	en	seo	google	untracked	Web
4ft3gnwmtx	2010-09-28	2.009061e+13	2010-08-02	FEMALE	56	basic	3	en	direct	direct	untracked	Web
bjjt8pjhuk	2011-12-05	2.009103e+13	2012-09-08	FEMALE	42	facebook	0	en	direct	direct	untracked	Web
87mebub9p4	2010-09-14	2.009121e+13	2010-02-18	-unknown-	41	basic	0	en	direct	direct	untracked	Web
osr2jwljor	2010-01-01	2.010010e+13	2010-01-02	-unknown-	NA	basic	0	en	other	other	omg	Web

Creating Account & First Booking Request

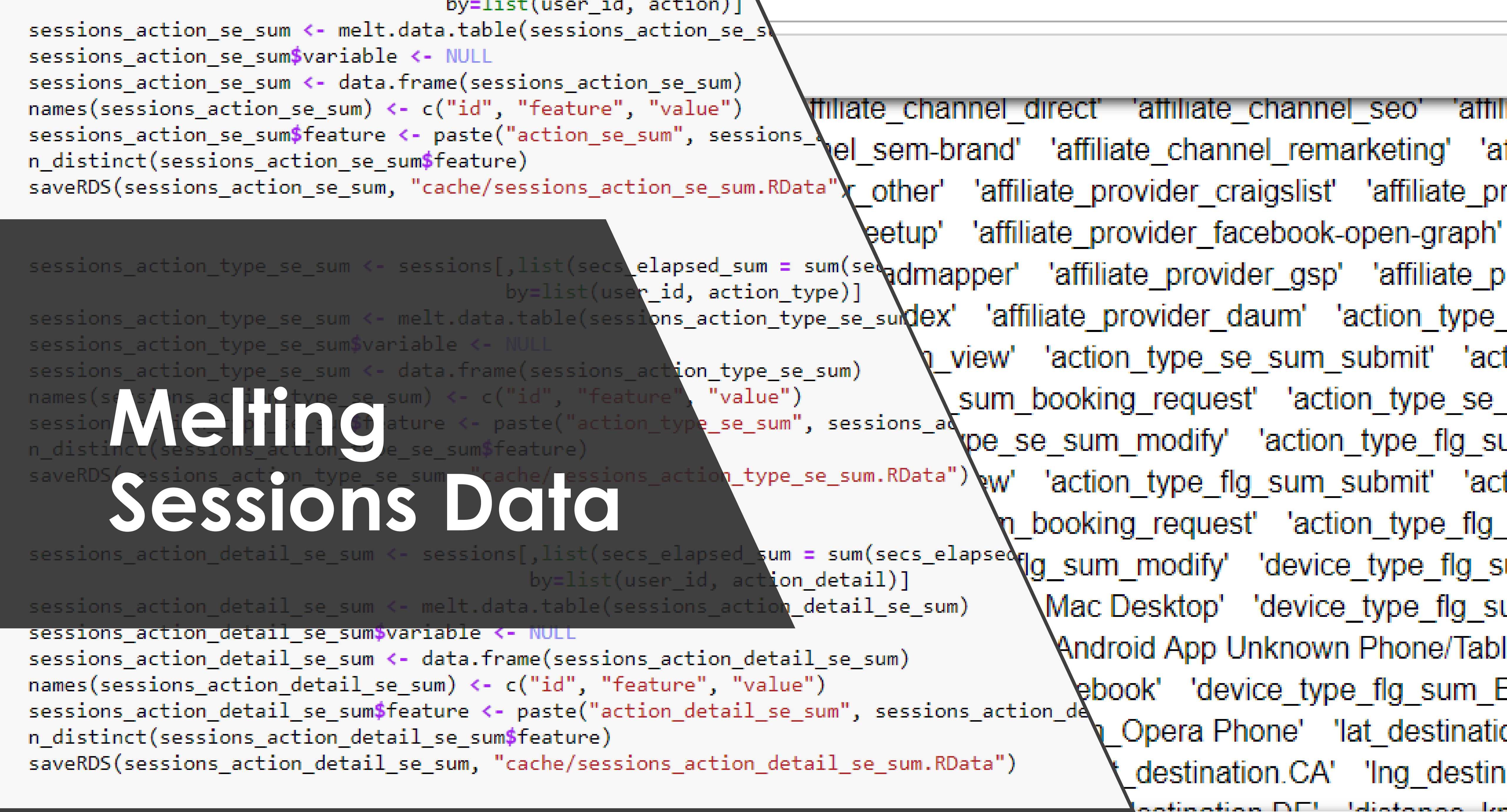


User Sessions Data



	user_id	action	action_type	action_detail	device_type	secs_elapsed	flg	seq	seq_rev	action2
108	d1mm9tcy42	show	view	p3	Windows Desktop	44446	1	108	20	show_view_p3_Windows Desktop
109	d1mm9tcy42	lookup			Windows Desktop	526	1	109	19	lookup___Windows Desktop
110	d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	11092	1	110	18	search_results_click_view_search_results_Windows Deskt...
111	d1mm9tcy42	lookup			Windows Desktop	702	1	111	17	lookup___Windows Desktop
112	d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	15435	1	112	16	search_results_click_view_search_results_Windows Deskt...
113	d1mm9tcy42	lookup			Windows Desktop	346	1	113	15	lookup___Windows Desktop
114	d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	11974	1	114	14	search_results_click_view_search_results_Windows Deskt...
115	d1mm9tcy42	personalize	data	wishlist_content_update	Windows Desktop	95	1	115	13	personalize_data_wishlist_content_update_Windows De...
116	d1mm9tcy42	lookup			Windows Desktop	2278	1	116	12	lookup___Windows Desktop
117	d1mm9tcy42	index	view	view_search_results	Windows Desktop	75484	1	117	11	index_view_view_search_results_Windows Desktop
118	d1mm9tcy42	decision_tree	-unknown-	-unknown-	-unknown-	89	1	118	10	decision_tree_-unknown-_-unknown-_-unknown-
119	d1mm9tcy42	recent_reservations	-unknown-	-unknown-	-unknown-	104	1	119	9	recent_reservations_-unknown-_-unknown-_-unknown-
120	d1mm9tcy42	faq_experiment_ids	-unknown-	-unknown-	-unknown-	49	1	120	8	faq_experiment_ids_-unknown-_-unknown-_-unknown-
121	d1mm9tcy42	multi	-unknown-	-unknown-	-unknown-	92	1	121	7	multi_-unknown-_-unknown-_-unknown-
122	d1mm9tcy42	active	-unknown-	-unknown-	-unknown-	111375	1	122	6	active_-unknown-_-unknown-_-unknown-
123	d1mm9tcy42	similar_listings	data	similar_listings	Windows Desktop	137	1	123	5	similar_listings_data_similar_listings_Windows Desktop
124	d1mm9tcy42	ajax_refresh_subtotal	click	change_trip_characteristics	Windows Desktop	791	1	124	4	ajax_refresh_subtotal_click_change_trip_characteristics_...
125	d1mm9tcy42	personalize	data	wishlist_content_update	Windows Desktop	73	1	125	3	personalize_data_wishlist_content_update_Windows De...
126	d1mm9tcy42	show			Windows Desktop	947	1	126	2	show___Windows Desktop
127	d1mm9tcy42	show	view	p3	Windows Desktop	76511	1	127	1	show_view_p3_Windows Desktop
128	yo8nz8bqcq	dashboard	view	dashboard	Mac Desktop	2739	1	1	9	dashboard_view_dashboard_Mac Desktop
129	yo8nz8bqcq	create	submit	create_user	Mac Desktop	NA	1	2	8	create_submit_create_user_Mac Desktop
130	yo8nz8bqcq	confirm_email	click	confirm_email_link	Mac Desktop	115983	1	3	7	confirm_email_click_confirm_email_link_Mac Desktop
131	yo8nz8bqcq	show	view	p3	Mac Desktop	20285	1	4	6	show_view_p3_Mac Desktop
132	yo8nz8bqcq	show_personalize	data	user_profile_content_update	Mac Desktop	3255	1	5	5	show_personalize_data_user_profile_content_update_M...
133	yo8nz8bqcq	show	view	user_profile	Mac Desktop	47308	1	6	4	show_view_user_profile_Mac Desktop
134	yo8nz8bqcq	header_userpic	data	header_userpic	Mac Desktop	14156	1	7	3	header_userpic_data_header_userpic_Mac Desktop
135	yo8nz8bqcq	personalize	data	wishlist_content_update	Mac Desktop	[illegible]	[illegible]	[illegible]	2	personalize_data_wishlist_content_update_Mac Desktop
136	yo8nz8bqcq	show			Mac Desktop	4080	1	9	1	show___Mac Desktop
137	4grx6yxeby	verify	-unknown-	-unknown-	Windows Desktop	65080	1	1	16	verify_-unknown-_-unknown-_Windows Desktop
138	4grx6yxeby	create	submit	create_user	Windows Desktop	NA	1	2	15	create_submit_create_user_Windows Desktop
139	4grx6yxeby		message_post	message_post	Windows Desktop	59801	1	3	14	_message_post_message_post_Windows Desktop

Melting Sessions Data

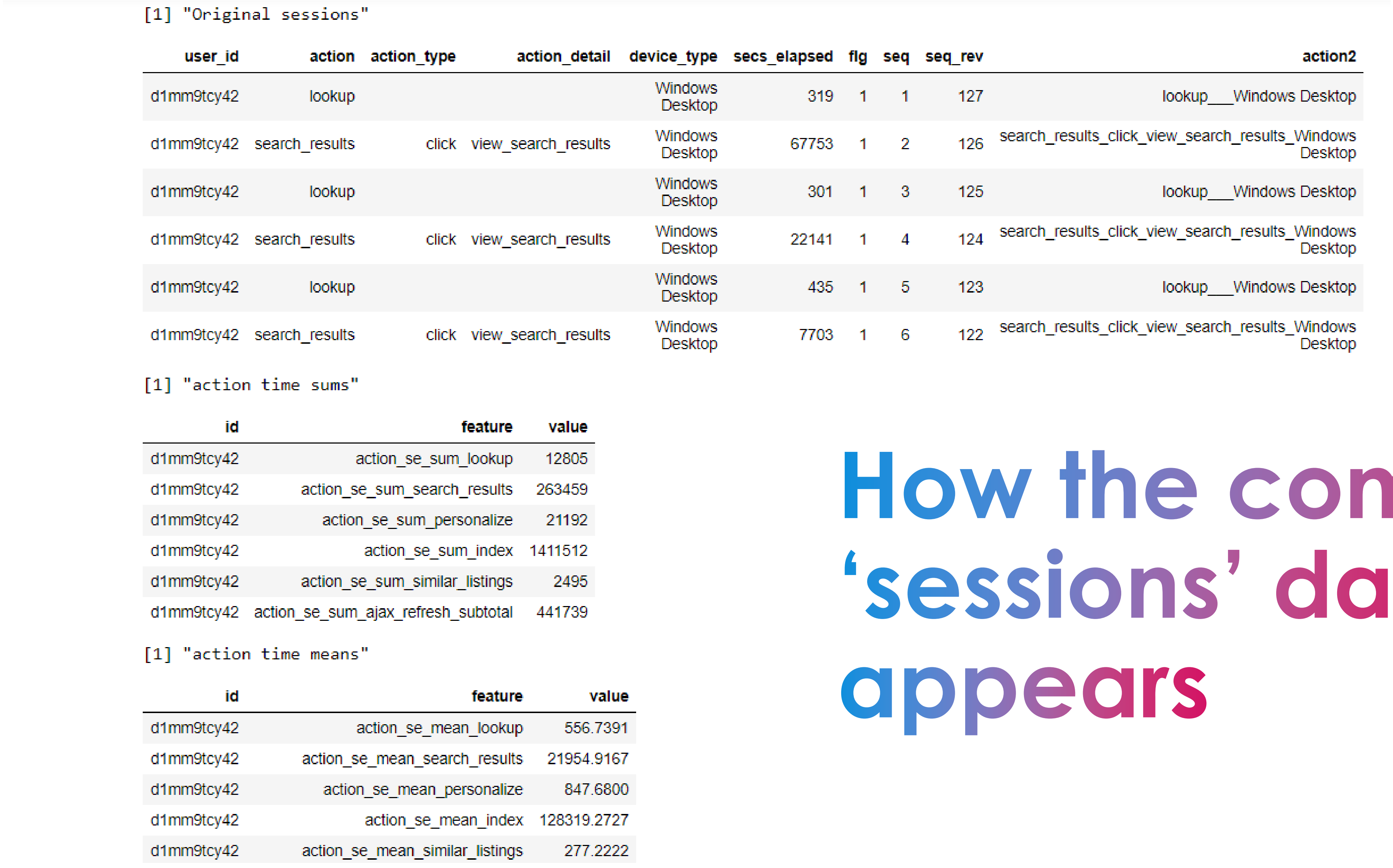


```
                                    by=list(user_id, action)]
sessions_action_se_sum <- melt.data.table(sessions_action_se_s
sessions_action_se_sum$variable <- NULL
sessions_action_se_sum <- data.frame(sessions_action_se_sum)
names(sessions_action_se_sum) <- c("id", "feature", "value")
sessions_action_se_sum$feature <- paste("action_se_sum", sessions_
n_distinct(sessions_action_se_sum$feature)
saveRDS(sessions_action_se_sum, "cache/sessions_action_se_sum.RData"

sessions_action_type_se_sum <- sessions[,list(secs_elapsed_sum = sum(se
                                         by=list(user_id, action_type)]
sessions_action_type_se_sum <- melt.data.table(sessions_action_type_se_su
sessions_action_type_se_sum$variable <- NULL
sessions_action_type_se_sum <- data.frame(sessions_action_type_se_sum)
names(sessions_action_type_se_sum) <- c("id", "feature", "value")
sessions_action_type_se_sum$feature <- paste("action_type_se_sum", sessions_ac
n_distinct(sessions_action_type_se_sum$feature)
saveRDS(sessions_action_type_se_sum, "cache/sessions_action_type_se_sum.RData")

sessions_action_detail_se_sum <- sessions[,list(secs_elapsed_sum = sum(secs_elapsed
                                           by=list(user_id, action_detail)]
sessions_action_detail_se_sum <- melt.data.table(sessions_action_detail_se_sum)
sessions_action_detail_se_sum$variable <- NULL
sessions_action_detail_se_sum <- data.frame(sessions_action_detail_se_sum)
names(sessions_action_detail_se_sum) <- c("id", "feature", "value")
sessions_action_detail_se_sum$feature <- paste("action_detail_se_sum", sessions_action_de
n_distinct(sessions_action_detail_se_sum$feature)
saveRDS(sessions_action_detail_se_sum, "cache/sessions_action_detail_se_sum.RData")
```

```
ffiliate_channel_direct'  'affiliate_channel_seo'  'affil
el_sem-brand'  'affiliate_channel_remarketing'  'af
r_other'  'affiliate_provider_craigslist'  'affiliate_pr
eetup'  'affiliate_provider_facebook-open-graph'
admapper'  'affiliate_provider_gsp'  'affiliate_p
dex'  'affiliate_provider_daum'  'action_type_
n_view'  'action_type_se_sum_submit'  'act
_sum_booking_request'  'action_type_se_
pe_se_sum_modify'  'action_type_flg_su
ew'  'action_type_flg_sum_submit'  'act
n_booking_request'  'action_type_flg_
lg_sum_modify'  'device_type_flg_s
Mac Desktop'  'device_type_flg_su
Android App Unknown Phone/Tabl
ebook'  'device_type_flg_sum_E
_Opera Phone'  'lat_destinatio
_destination.CA'  'lng_destin
```



```
[1] "Original sessions"
```

user_id	action	action_type	action_detail	device_type	secs_elapsed	flg	seq	seq_rev	action2
d1mm9tcy42	lookup			Windows Desktop	319	1	1	127	lookup___Windows Desktop
d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	67753	1	2	126	search_results_click_view_search_results_Windows Desktop
d1mm9tcy42	lookup			Windows Desktop	301	1	3	125	lookup___Windows Desktop
d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	22141	1	4	124	search_results_click_view_search_results_Windows Desktop
d1mm9tcy42	lookup			Windows Desktop	435	1	5	123	lookup___Windows Desktop
d1mm9tcy42	search_results	click	view_search_results	Windows Desktop	7703	1	6	122	search_results_click_view_search_results_Windows Desktop

```
[1] "action time sums"
```

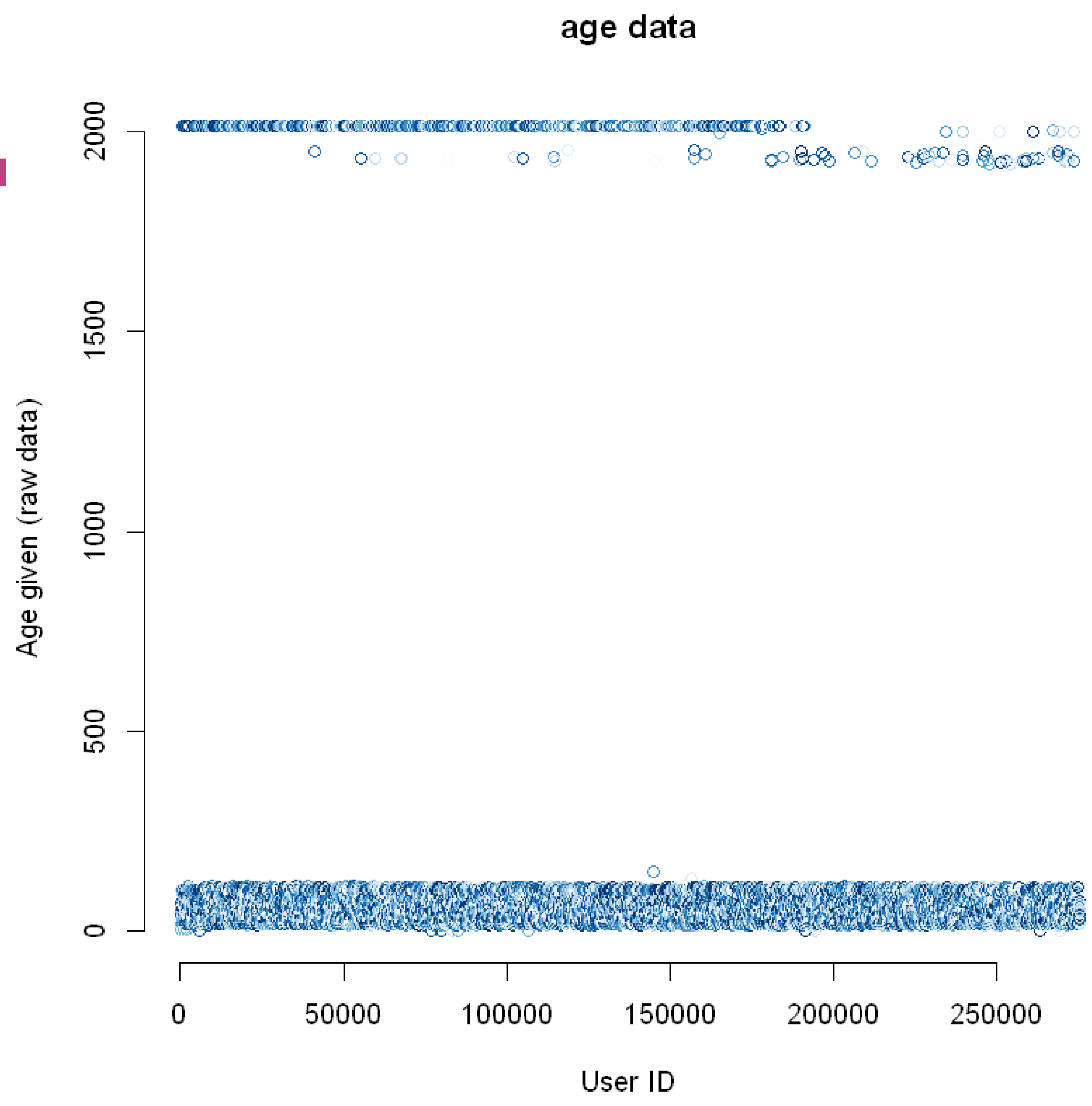
id	feature	value
d1mm9tcy42	action_se_sum_lookup	12805
d1mm9tcy42	action_se_sum_search_results	263459
d1mm9tcy42	action_se_sum_personalize	21192
d1mm9tcy42	action_se_sum_index	1411512
d1mm9tcy42	action_se_sum_similar_listings	2495
d1mm9tcy42	action_se_sum_ajax_refresh_subtotal	441739

```
[1] "action time means"
```

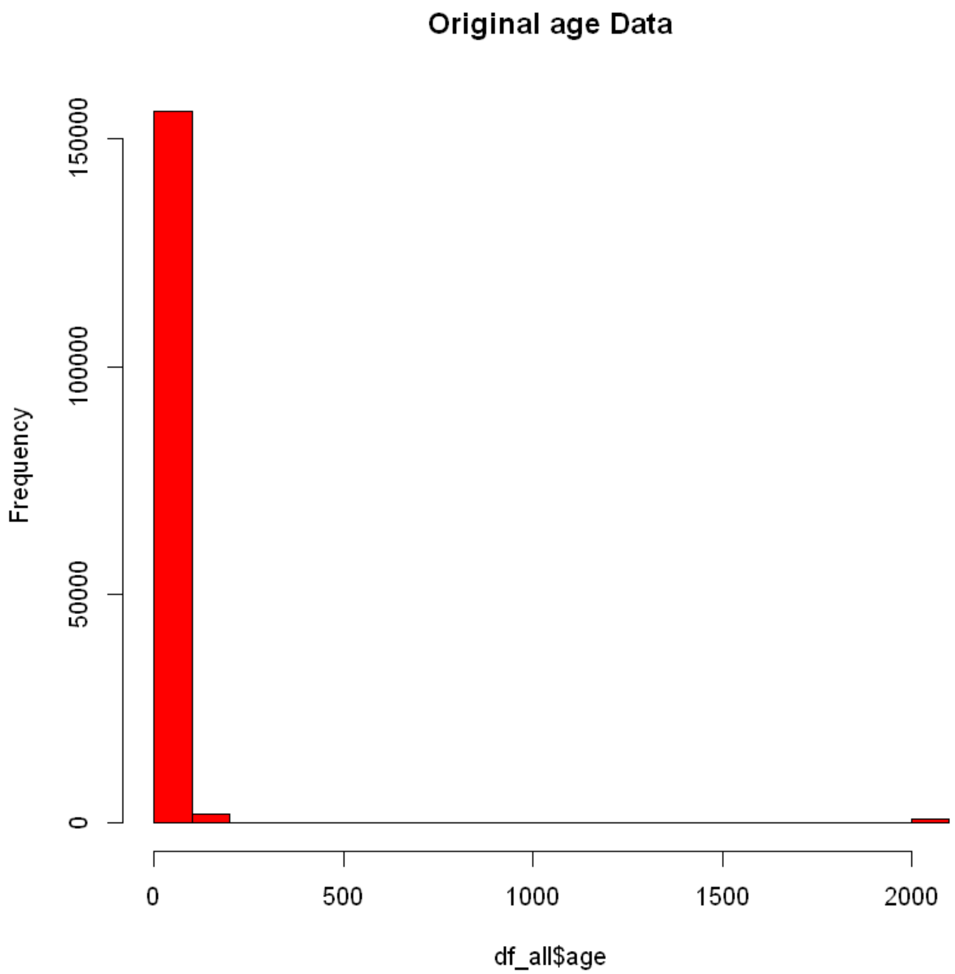
id	feature	value
d1mm9tcy42	action_se_mean_lookup	556.7391
d1mm9tcy42	action_se_mean_search_results	21954.9167
d1mm9tcy42	action_se_mean_personalize	847.6800
d1mm9tcy42	action_se_mean_index	128319.2727
d1mm9tcy42	action_se_mean_similar_listings	277.2222

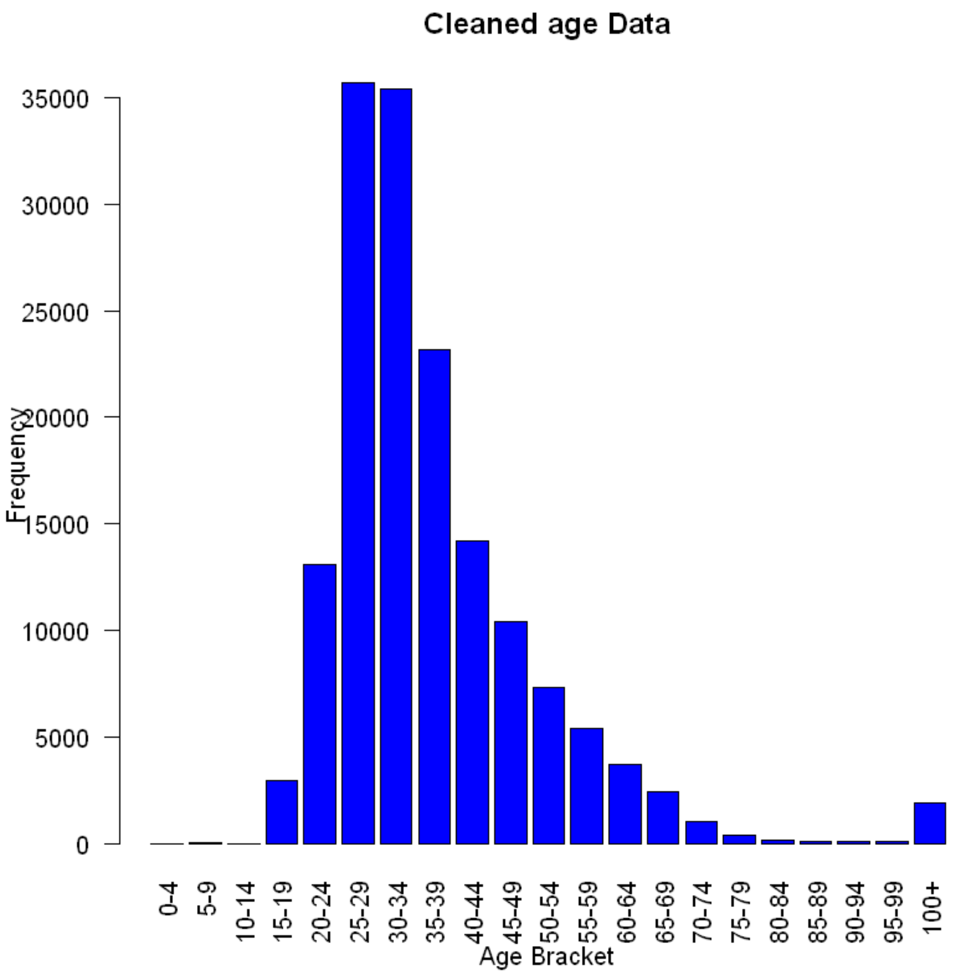
How the combined ‘sessions’ data appears

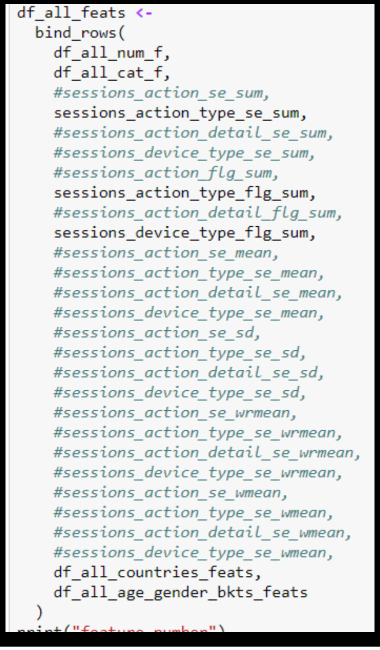
Age Data


age data
Age given (raw data)
0
500
1000
1500
2000
User ID
0
50000
100000
150000
200000
250000

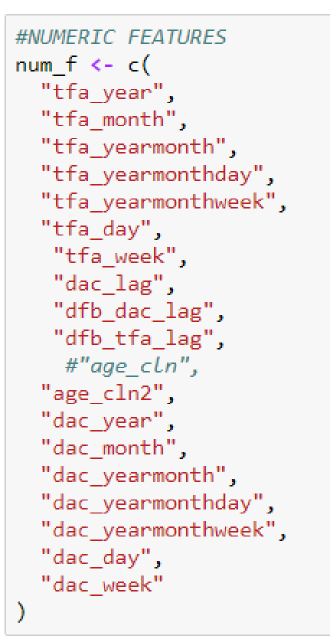
Age Brackets



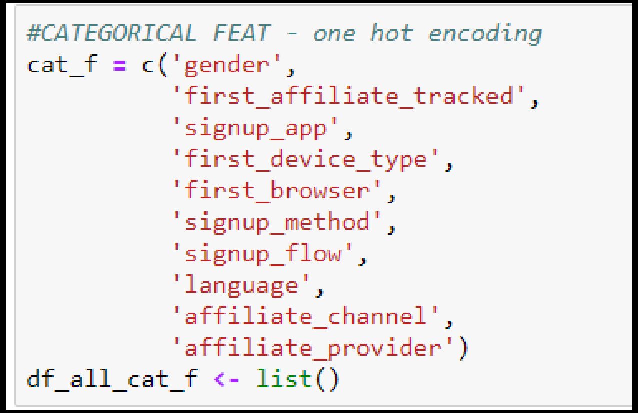




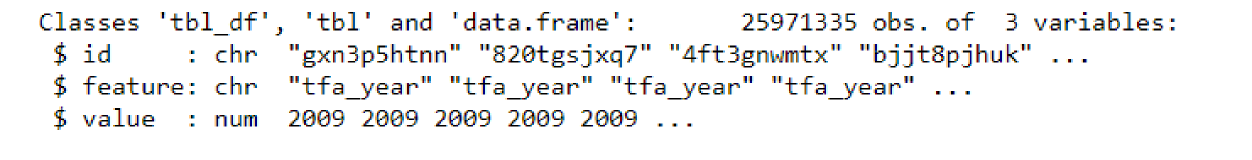
```
df_all_feats <-
  bind_rows(
    df_all_num_f,
    df_all_cat_f,
    #sessions_action_se_sum,
    sessions_action_type_se_sum,
    #sessions_action_detail_se_sum,
    #sessions_device_type_se_sum,
    #sessions_action_flg_sum,
    sessions_action_type_flg_sum,
    #sessions_action_detail_flg_sum,
    sessions_device_type_flg_sum,
    #sessions_action_se_mean,
    #sessions_action_type_se_mean,
    #sessions_action_detail_se_mean,
    #sessions_device_type_se_mean,
    #sessions_action_se_sd,
    #sessions_action_type_se_sd,
    #sessions_action_detail_se_sd,
    #sessions_device_type_se_sd,
    #sessions_action_se_wrmean,
    #sessions_action_type_se_wrmean,
    #sessions_action_detail_se_wrmean,
    #sessions_device_type_se_wrmean,
    #sessions_action_se_wmean,
    #sessions_action_type_se_wmean,
    #sessions_action_detail_se_wmean,
    #sessions_device_type_se_wmean,
    df_all_countries_feats,
    df_all_age_gender_bkts_feats
  )
print("feature number")
```



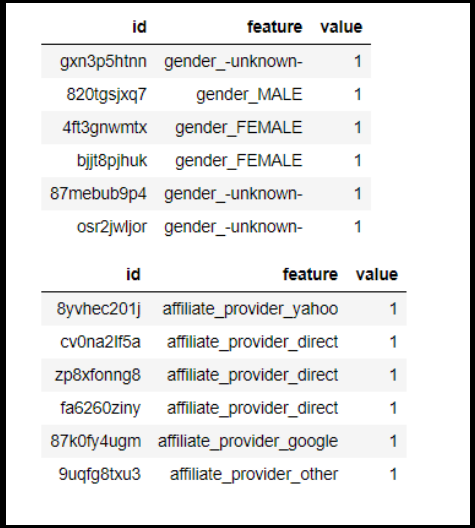
```
#NUMERIC FEATURES
num_f <- c(
  "tfa_year",
  "tfa_month",
  "tfa_yearmonth",
  "tfa_yearmonthday",
  "tfa_yearmonthweek",
  "tfa_day",
   "tfa_week",
   "dac_lag",
   "dfb_dac_lag",
   "dfb_tfa_lag",
    #"age_cln",
  "age_cln2",
  "dac_year",
  "dac_month",
  "dac_yearmonth",
  "dac_yearmonthday",
  "dac_yearmonthweek",
  "dac_day",
  "dac_week"
)
```



```
#CATEGORICAL FEAT - one hot encoding
cat_f = c('gender',
          'first_affiliate_tracked',
          'signup_app',
          'first_device_type',
          'first_browser',
          'signup_method',
          'signup_flow',
          'language',
          'affiliate_channel',
          'affiliate_provider')
df_all_cat_f <- list()
```



```
Classes 'tbl_df', 'tbl' and 'data.frame':       25971335 obs. of  3 variables:
 $ id     : chr  "gxn3p5htnn" "820tgsjxq7" "4ft3gnwmtx" "bjjt8pjhuk" ...
 $ feature: chr  "tfa_year" "tfa_year" "tfa_year" "tfa_year" ...
 $ value  : num  2009 2009 2009 2009 2009 ...
```



id	feature	value
gxn3p5htnn	gender_-unknown-	1
820tgsjxq7	gender_MALE	1
4ft3gnwmtx	gender_FEMALE	1
bjjt8pjhuk	gender_FEMALE	1
87mebub9p4	gender_-unknown-	1
osr2jwljor	gender_-unknown-	1

id	feature	value
8yvhec201j	affiliate_provider_yahoo	1
cv0na2lf5a	affiliate_provider_direct	1
zp8xfonng8	affiliate_provider_direct	1
fa6260ziny	affiliate_provider_direct	1
87k0fy4ugm	affiliate_provider_google	1
9uqfg8txu3	affiliate_provider_other	1

XGBoost

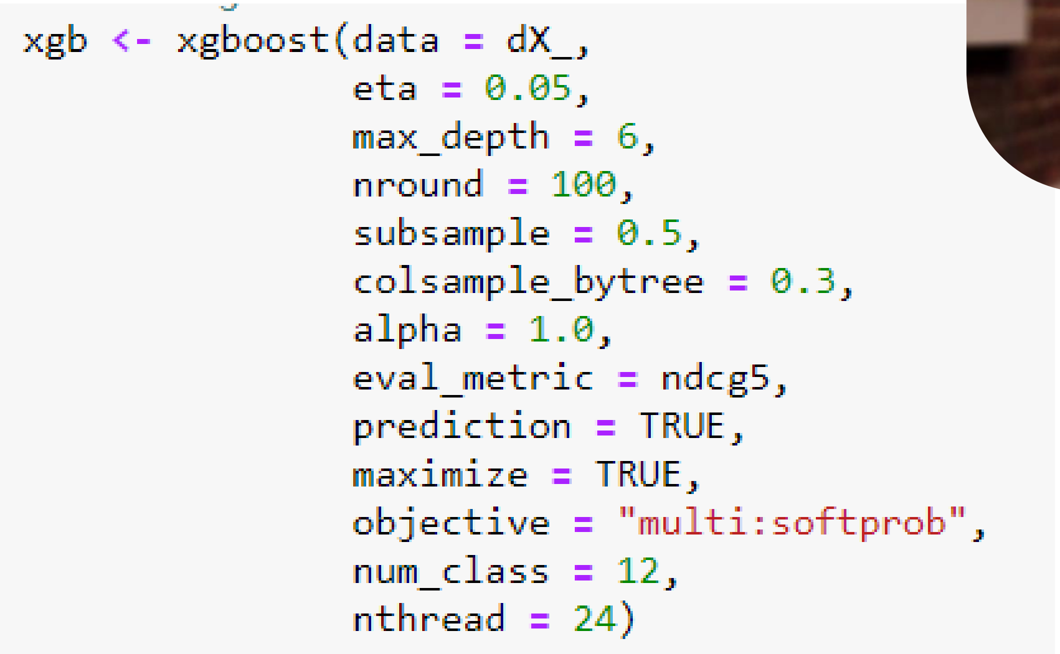
Train: 193116 x 269

Val: 20335 x 269

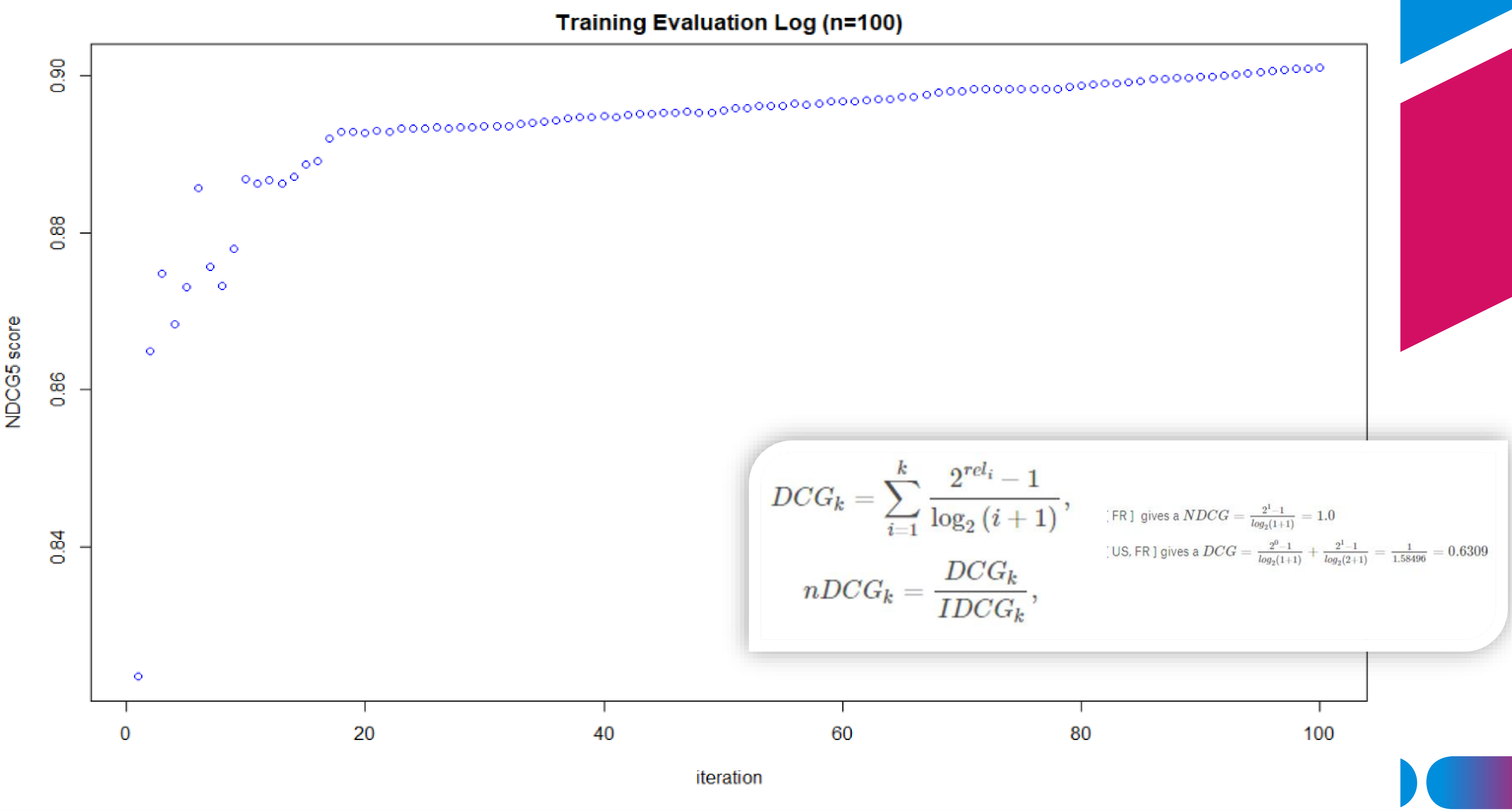
Test: 62096 x 269

Train/Val : xgb.cv()

Train:



```
xgb <- xgboost(data = dX_,
               eta = 0.05,
               max_depth = 6,
               nround = 100,
               subsample = 0.5,
               colsample_bytree = 0.3,
               alpha = 1.0,
               eval_metric = ndcg5,
               prediction = TRUE,
               maximize = TRUE,
               objective = "multi:softprob",
               num_class = 12,
               nthread = 24)
```



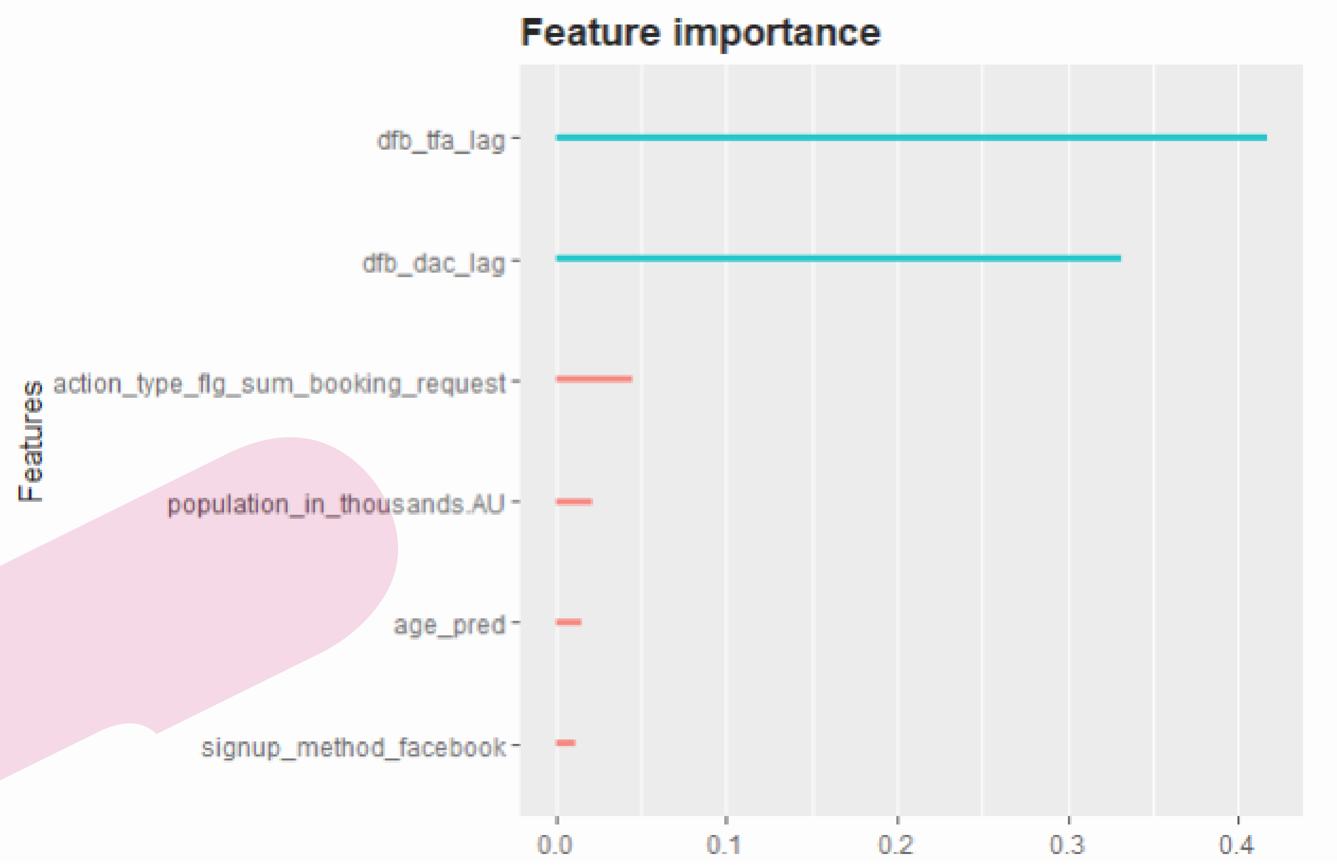
$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2(i+1)},$$

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

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

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

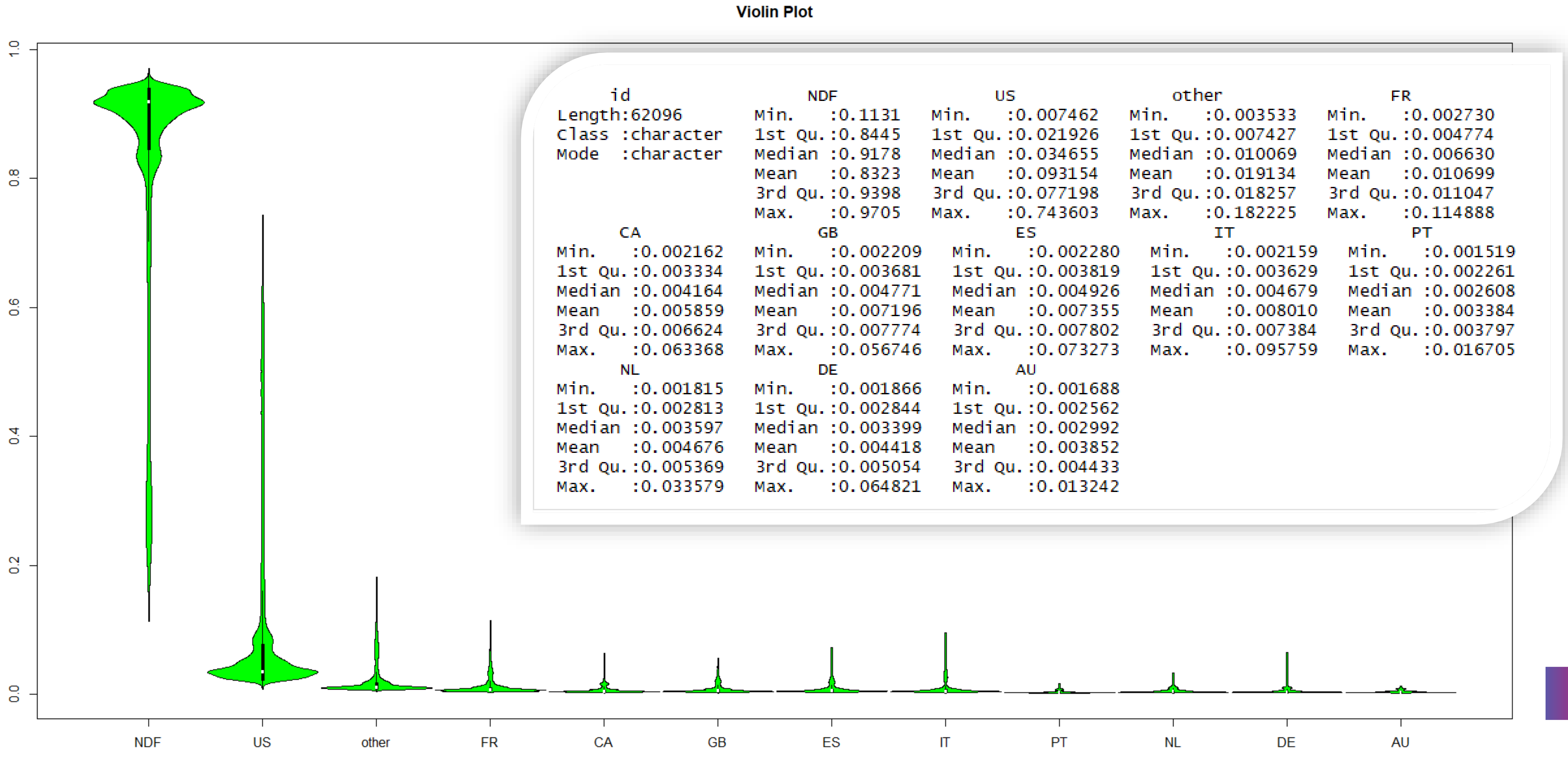
XGBOOST TRAINING LOG



Feature Importance

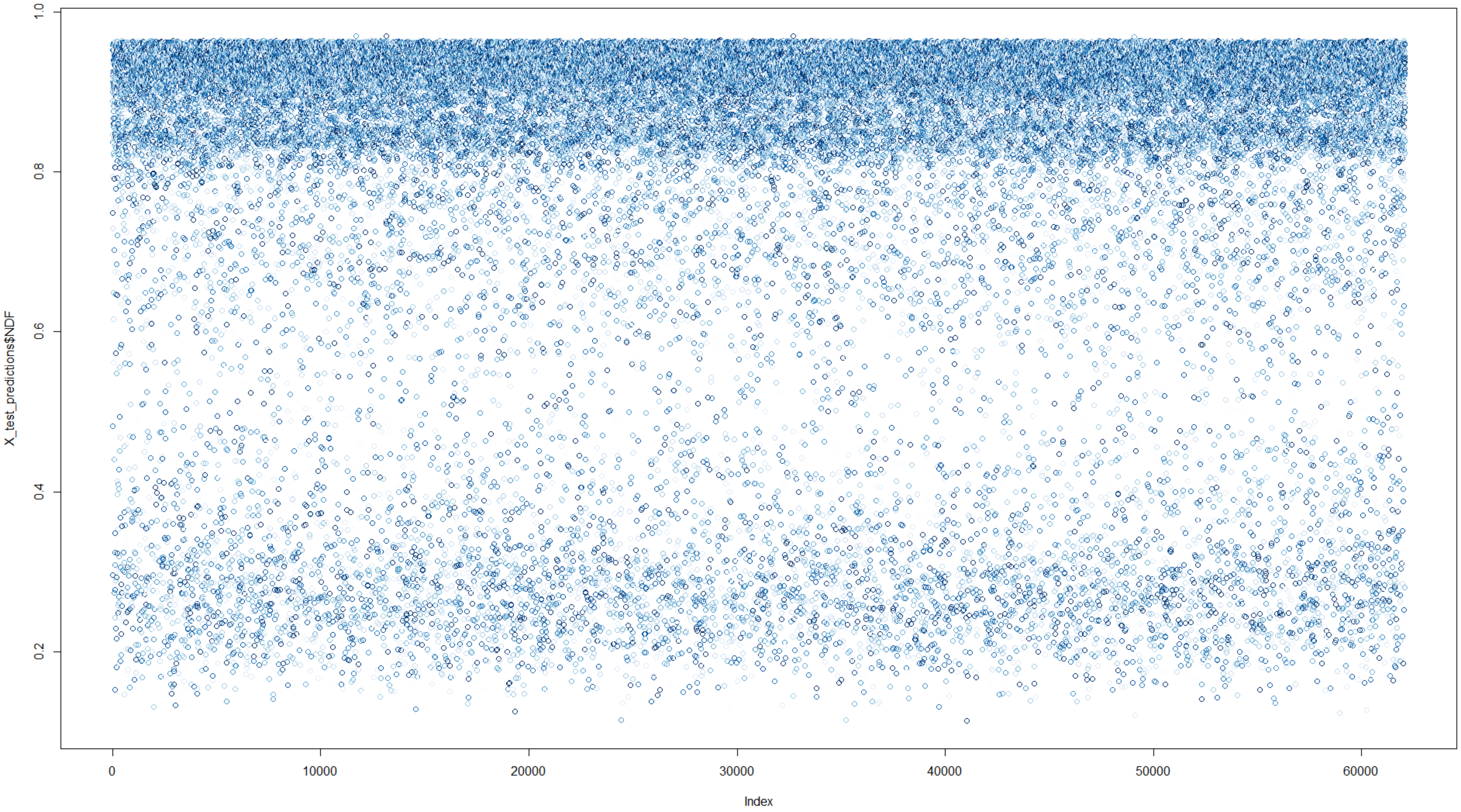
- Lag between booking request and first active
- Lag between booking request and account creation
- Existence of a booking request
- Stat of population of an age bracket
- Predicted Age
- Signing up with Facebook

Predicted Probability per Destination Class

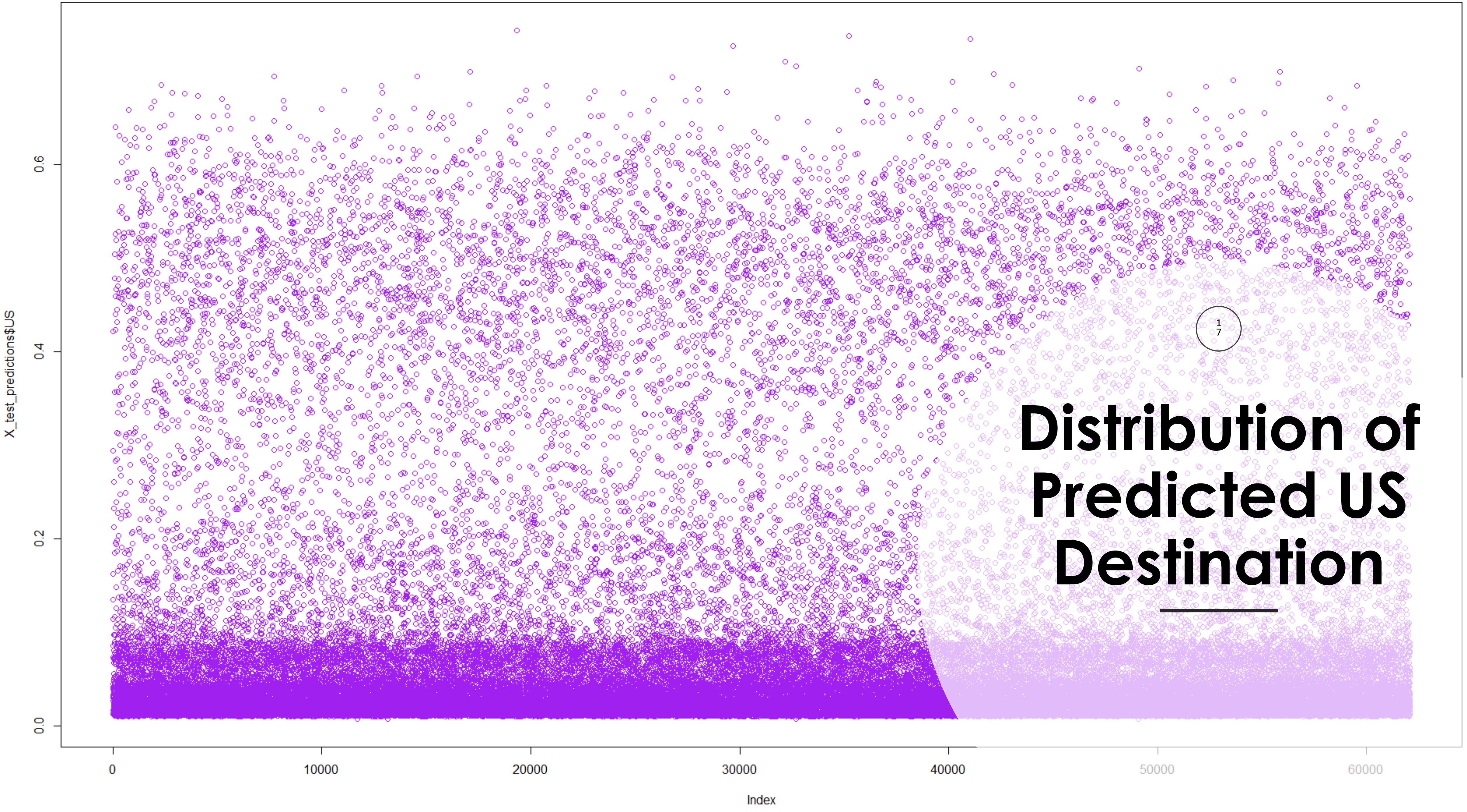


```
      id                 NDF                US                other                FR
Length:62096       Min.   :0.1131   Min.   :0.007462   Min.   :0.003533   Min.   :0.002730
Class :character   1st Qu.:0.8445   1st Qu.:0.021926   1st Qu.:0.007427   1st Qu.:0.004774
Mode  :character   Median :0.9178   Median :0.034655   Median :0.010069   Median :0.006630
                   Mean   :0.8323   Mean   :0.093154   Mean   :0.019134   Mean   :0.010699
                   3rd Qu.:0.9398   3rd Qu.:0.077198   3rd Qu.:0.018257   3rd Qu.:0.011047
                   Max.   :0.9705   Max.   :0.743603   Max.   :0.182225   Max.   :0.114888
       CA                  GB                  ES                  IT                  PT
Min.   :0.002162   Min.   :0.002209   Min.   :0.002280   Min.   :0.002159   Min.   :0.001519
1st Qu.:0.003334   1st Qu.:0.003681   1st Qu.:0.003819   1st Qu.:0.003629   1st Qu.:0.002261
Median :0.004164   Median :0.004771   Median :0.004926   Median :0.004679   Median :0.002608
Mean   :0.005859   Mean   :0.007196   Mean   :0.007355   Mean   :0.008010   Mean   :0.003384
3rd Qu.:0.006624   3rd Qu.:0.007774   3rd Qu.:0.007802   3rd Qu.:0.007384   3rd Qu.:0.003797
Max.   :0.063368   Max.   :0.056746   Max.   :0.073273   Max.   :0.095759   Max.   :0.016705
       NL                  DE                  AU
Min.   :0.001815   Min.   :0.001866   Min.   :0.001688
1st Qu.:0.002813   1st Qu.:0.002844   1st Qu.:0.002562
Median :0.003597   Median :0.003399   Median :0.002992
Mean   :0.004676   Mean   :0.004418   Mean   :0.003852
3rd Qu.:0.005369   3rd Qu.:0.005054   3rd Qu.:0.004433
Max.   :0.033579   Max.   :0.064821   Max.   :0.013242
```

Distribution of No Destination Found Predictions

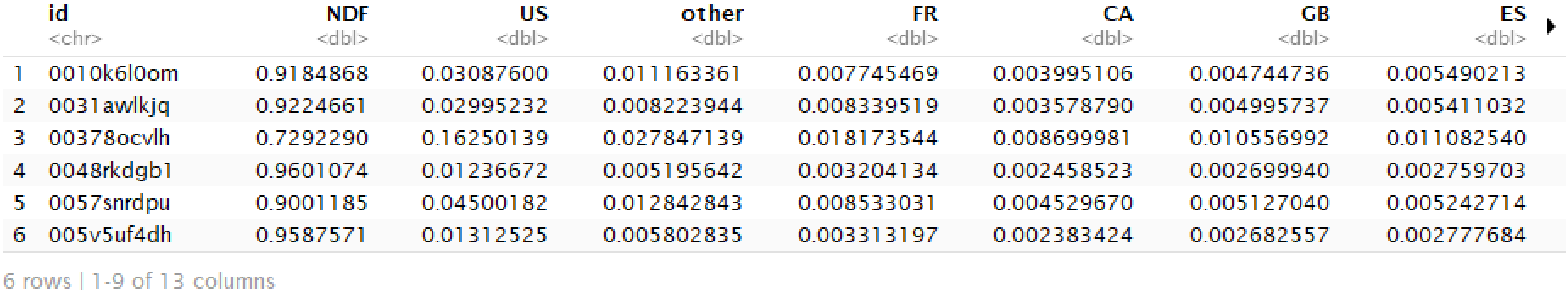


Distribution of Predicted US Destination



Formatting Solution

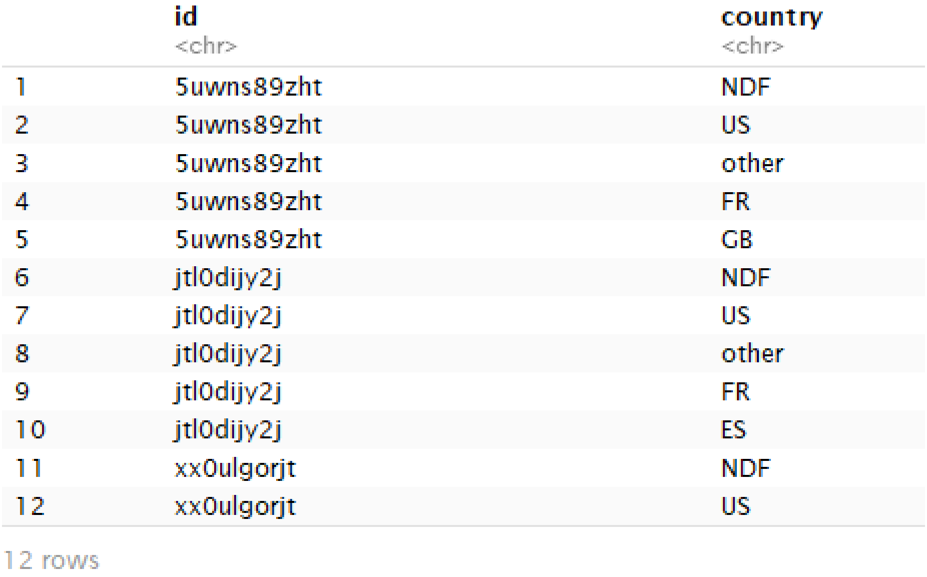
XGB output format



	id <chr>	NDF <dbl>	US <dbl>	other <dbl>	FR <dbl>	CA <dbl>	GB <dbl>	ES <dbl>
1	0010k6l0om	0.9184868	0.03087600	0.011163361	0.007745469	0.003995106	0.004744736	0.005490213
2	0031awlkjq	0.9224661	0.02995232	0.008223944	0.008339519	0.003578790	0.004995737	0.005411032
3	00378ocvlh	0.7292290	0.16250139	0.027847139	0.018173544	0.008699981	0.010556992	0.011082540
4	0048rkdgb1	0.9601074	0.01236672	0.005195642	0.003204134	0.002458523	0.002699940	0.002759703
5	0057snrdpu	0.9001185	0.04500182	0.012842843	0.008533031	0.004529670	0.005127040	0.005242714
6	005v5uf4dh	0.9587571	0.01312525	0.005802835	0.003313197	0.002383424	0.002682557	0.002777684

6 rows | 1-9 of 13 columns

Submission format



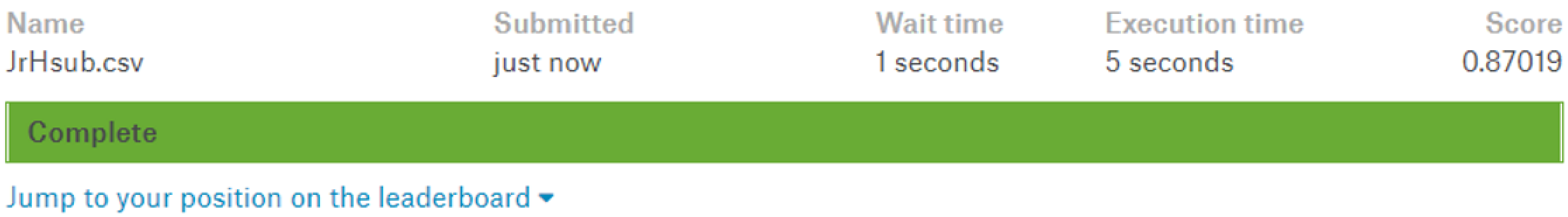
	id <chr>	country <chr>
1	5uwns89zht	NDF
2	5uwns89zht	US
3	5uwns89zht	other
4	5uwns89zht	FR
5	5uwns89zht	GB
6	jtl0dijy2j	NDF
7	jtl0dijy2j	US
8	jtl0dijy2j	other
9	jtl0dijy2j	FR
10	jtl0dijy2j	ES
11	xx0ulgorjt	NDF
12	xx0ulgorjt	US

12 rows

SUBMISSION

Scored #415 out of 1,462

70th Percentile



Name	Submitted	Wait time	Execution time	Score
JrHsub.csv	just now	1 seconds	5 seconds	0.87019

Complete

THANK YOU!

JEANETTE HENRY

Email:
Jeanette.Henry@live.com