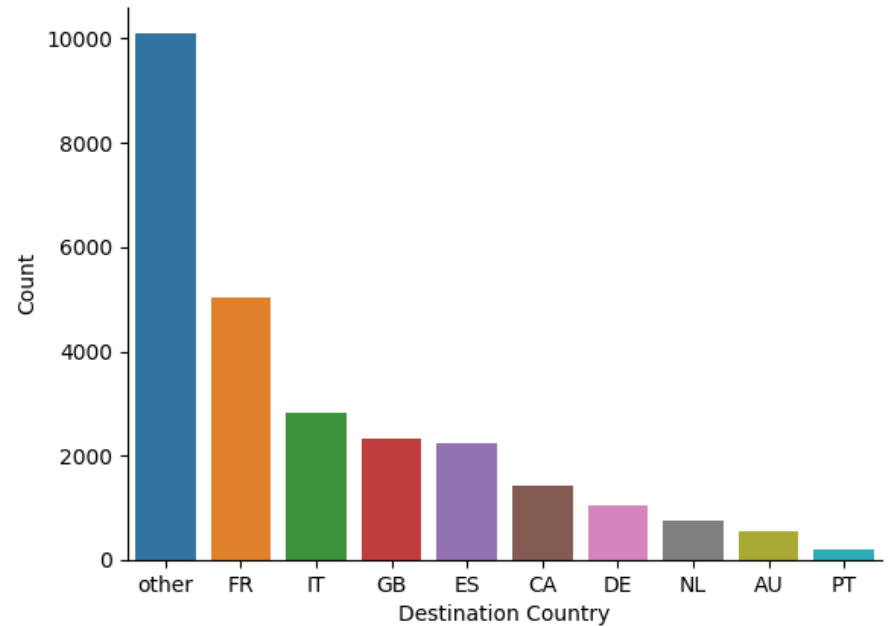
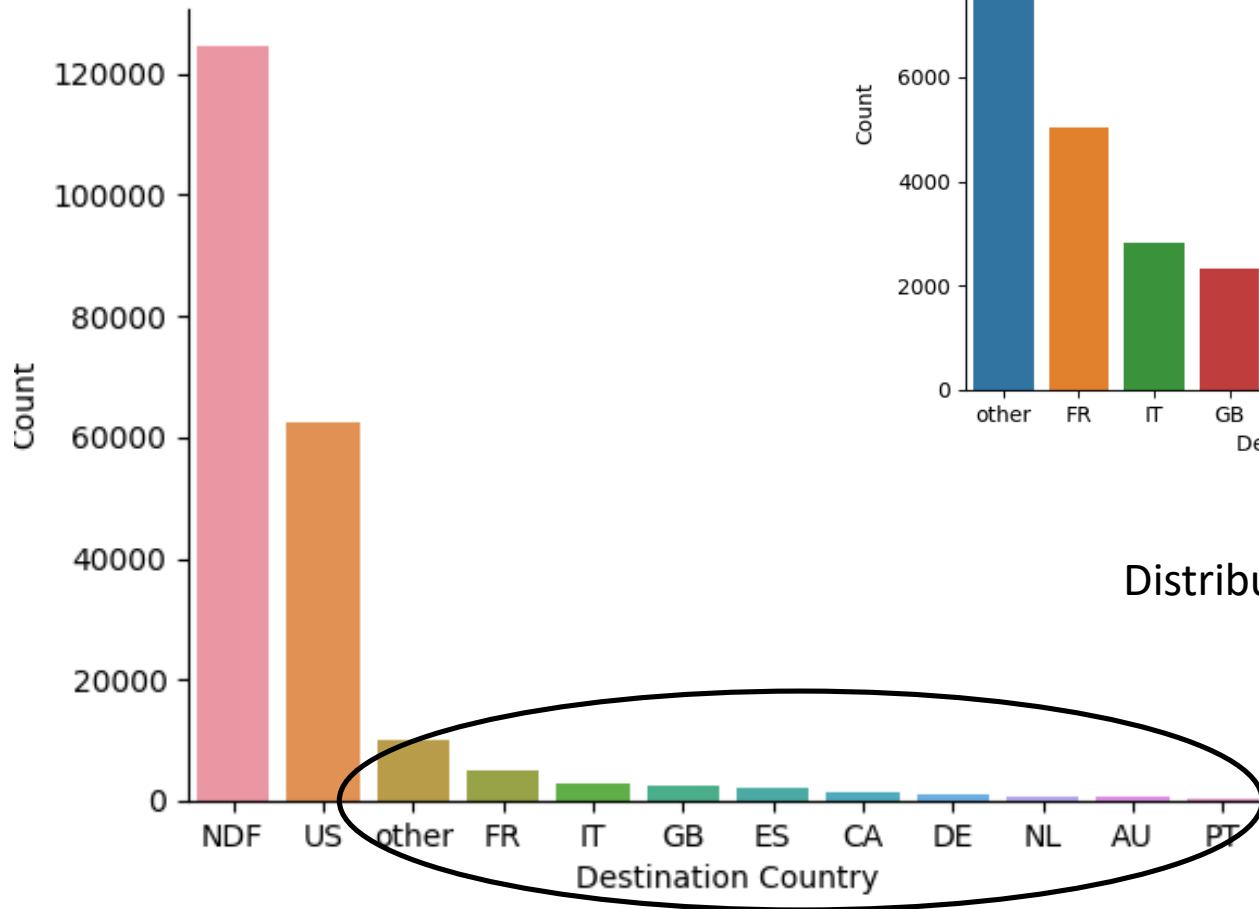


Airbnb new user bookings

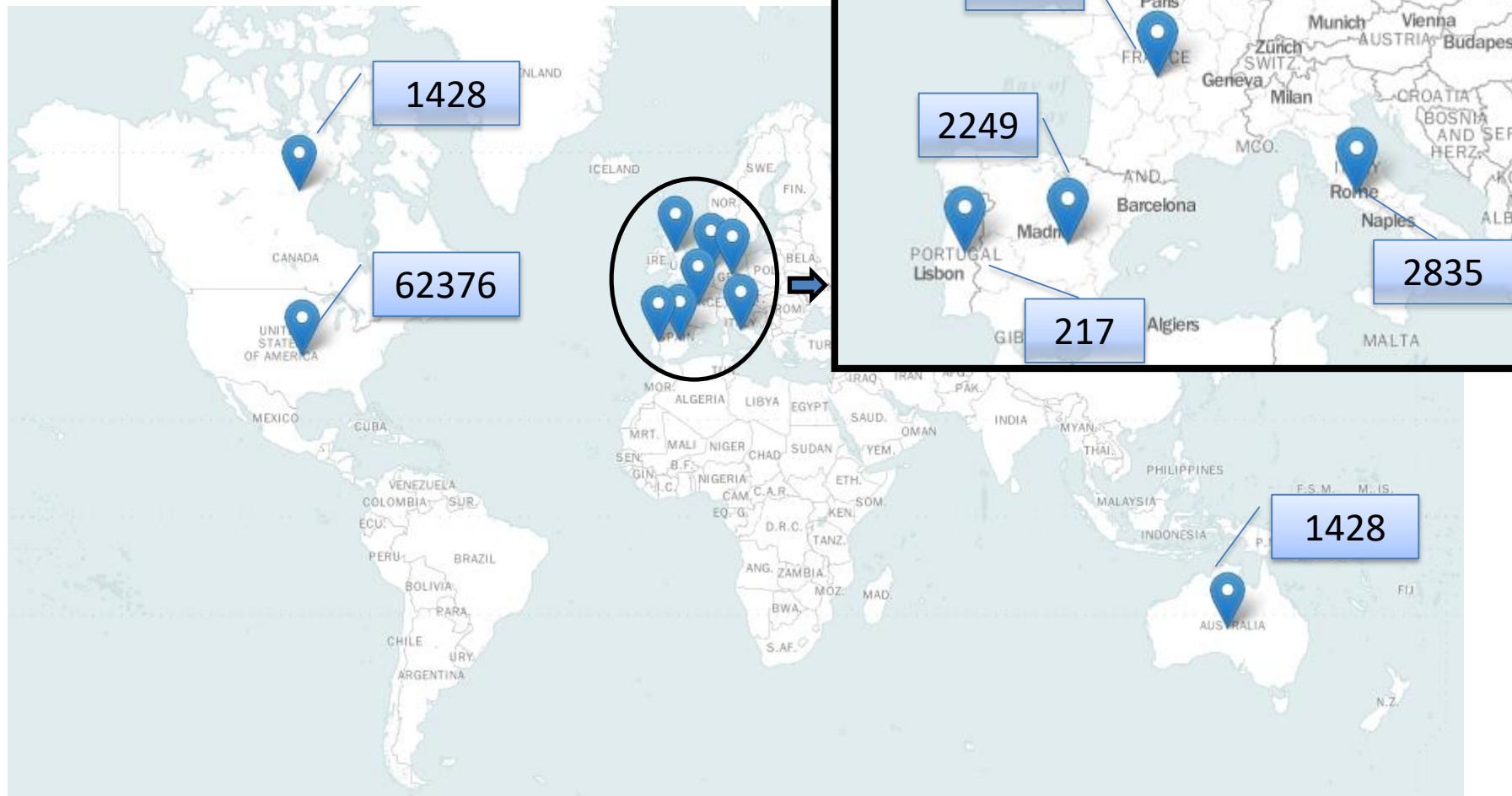
Group 8

Data Exploration

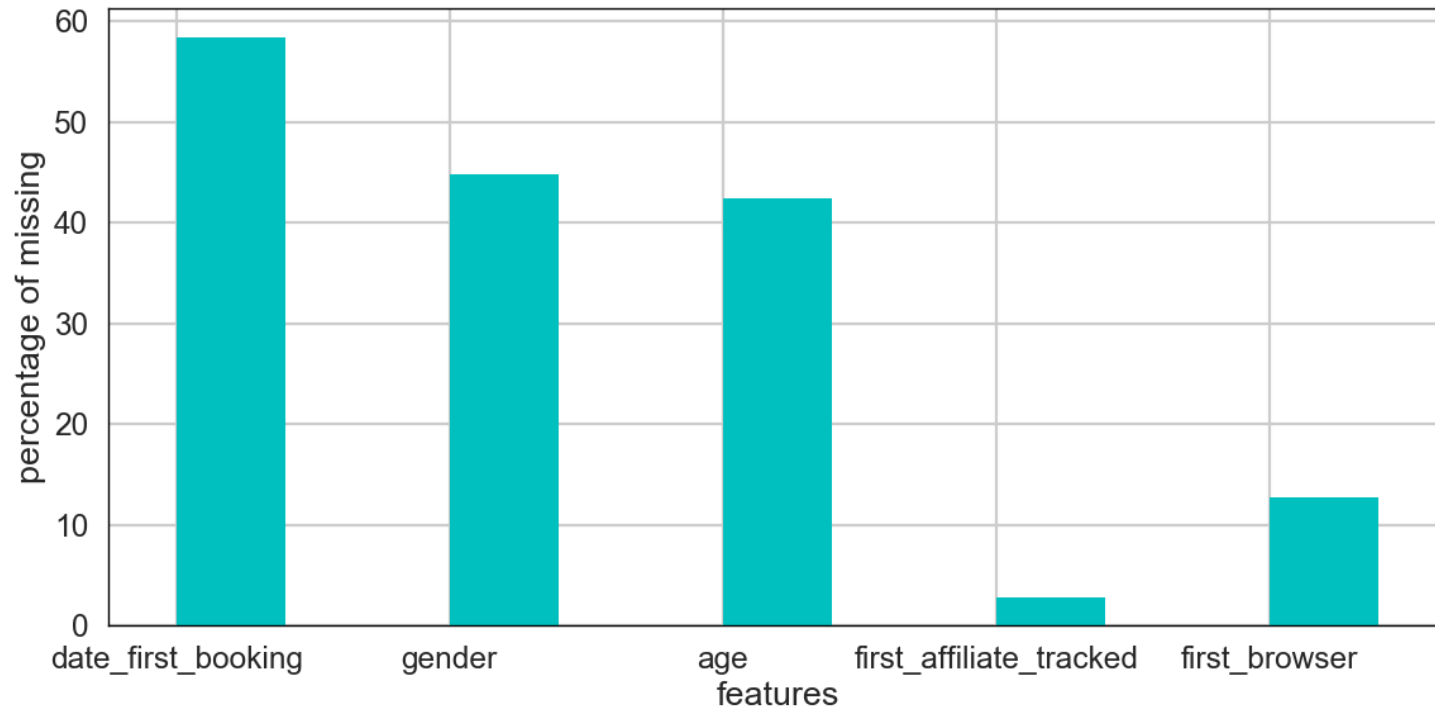


Distribution of destinations

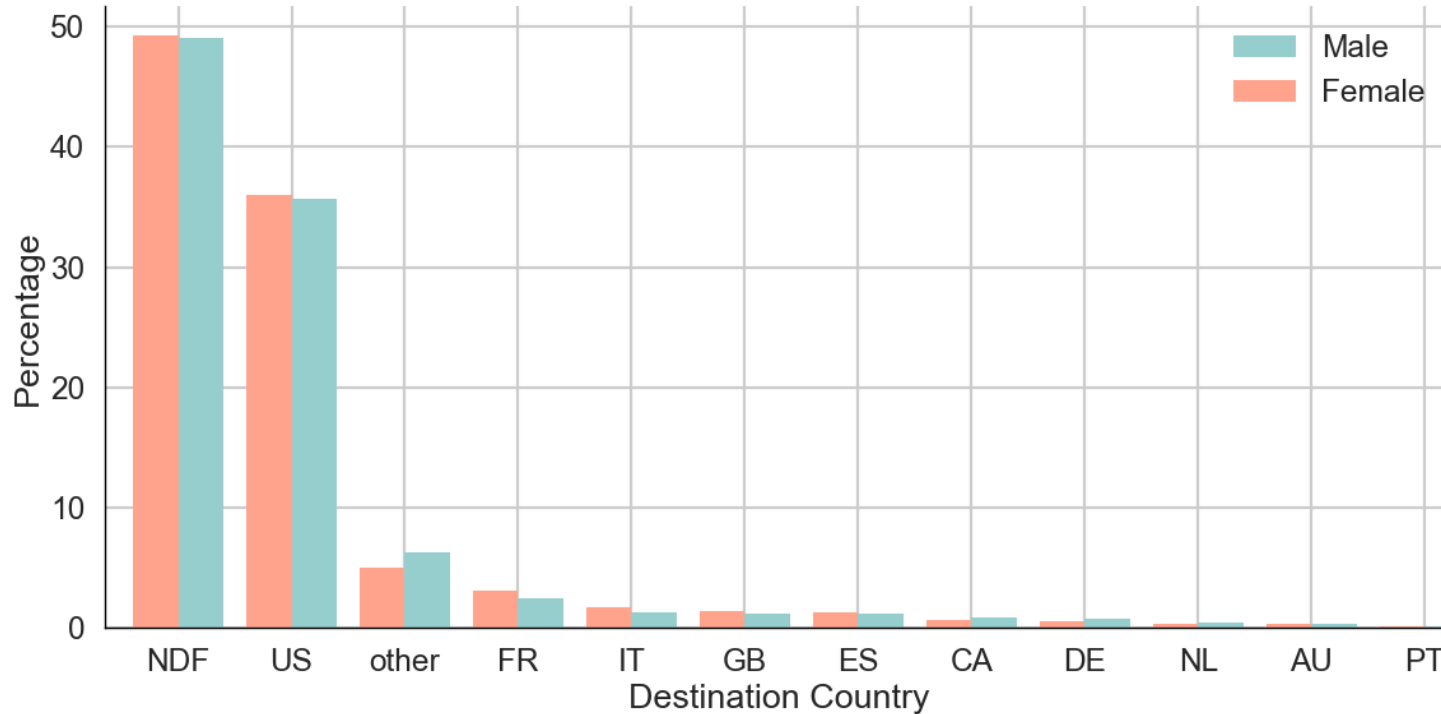
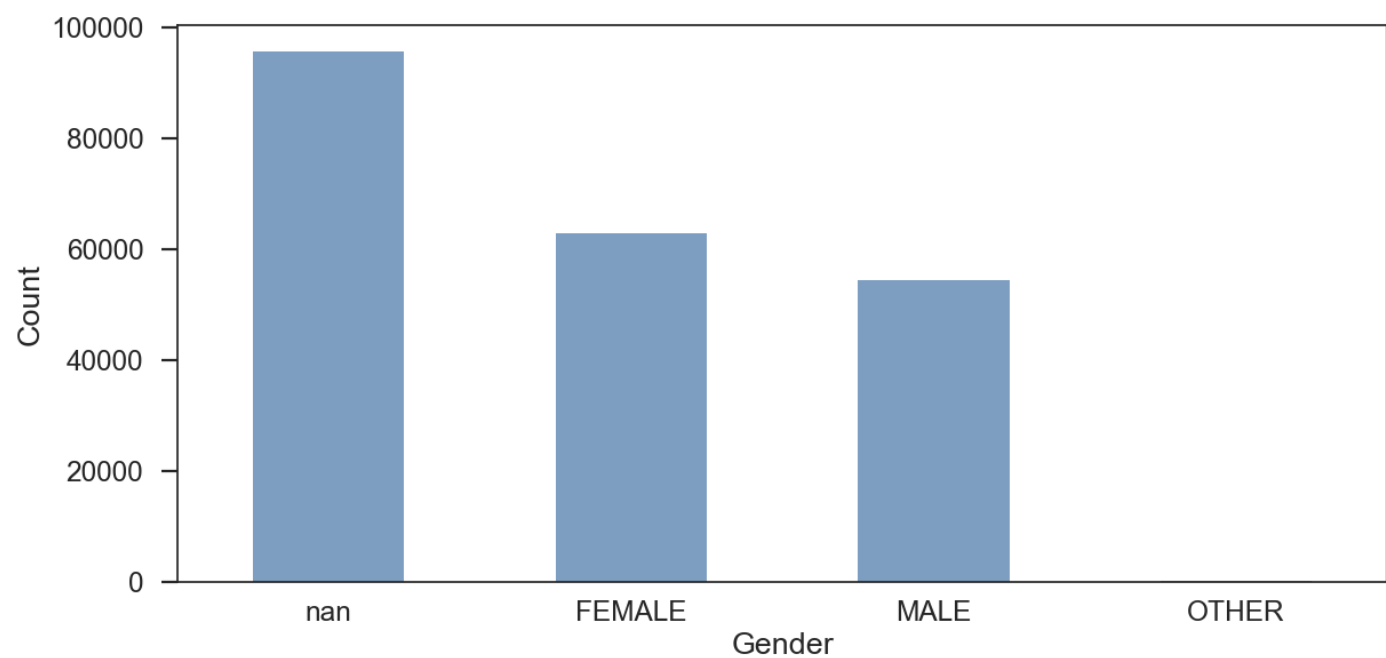
The destinations' location and their counts are also shown in the map with pins in the location and marked with total population of who choose it as first destination.



Features with missing value

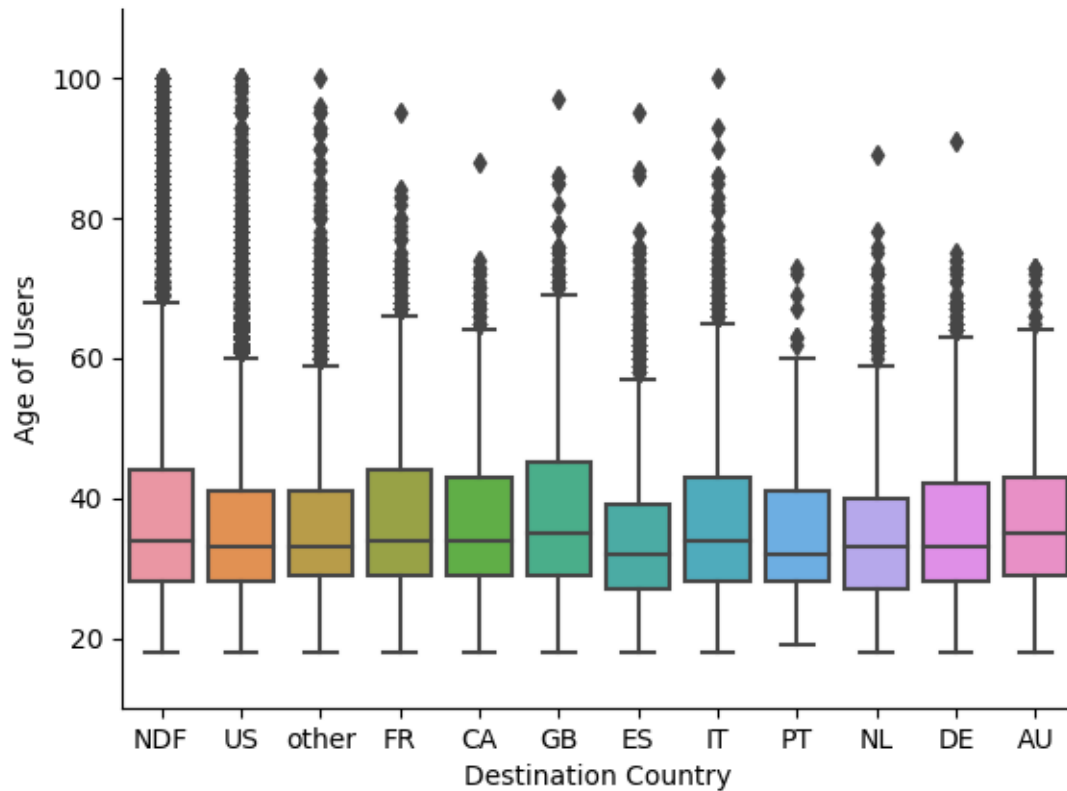
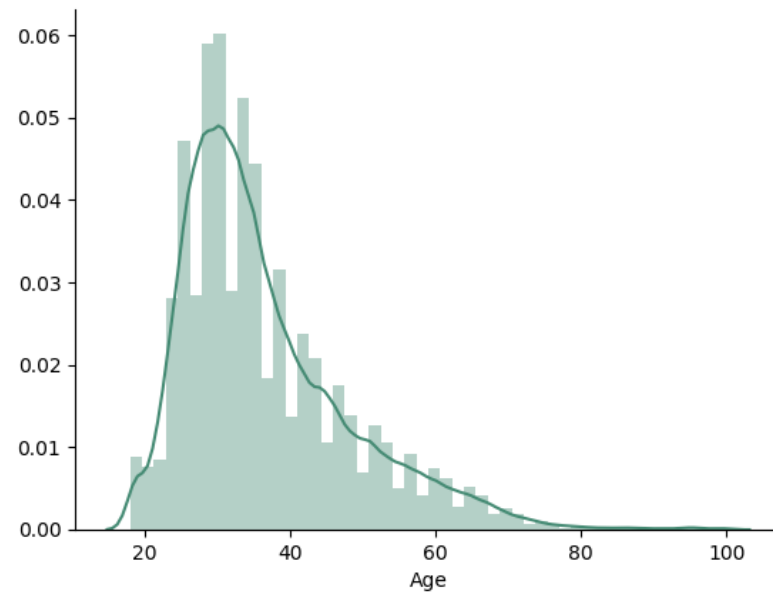


Gender
histogram



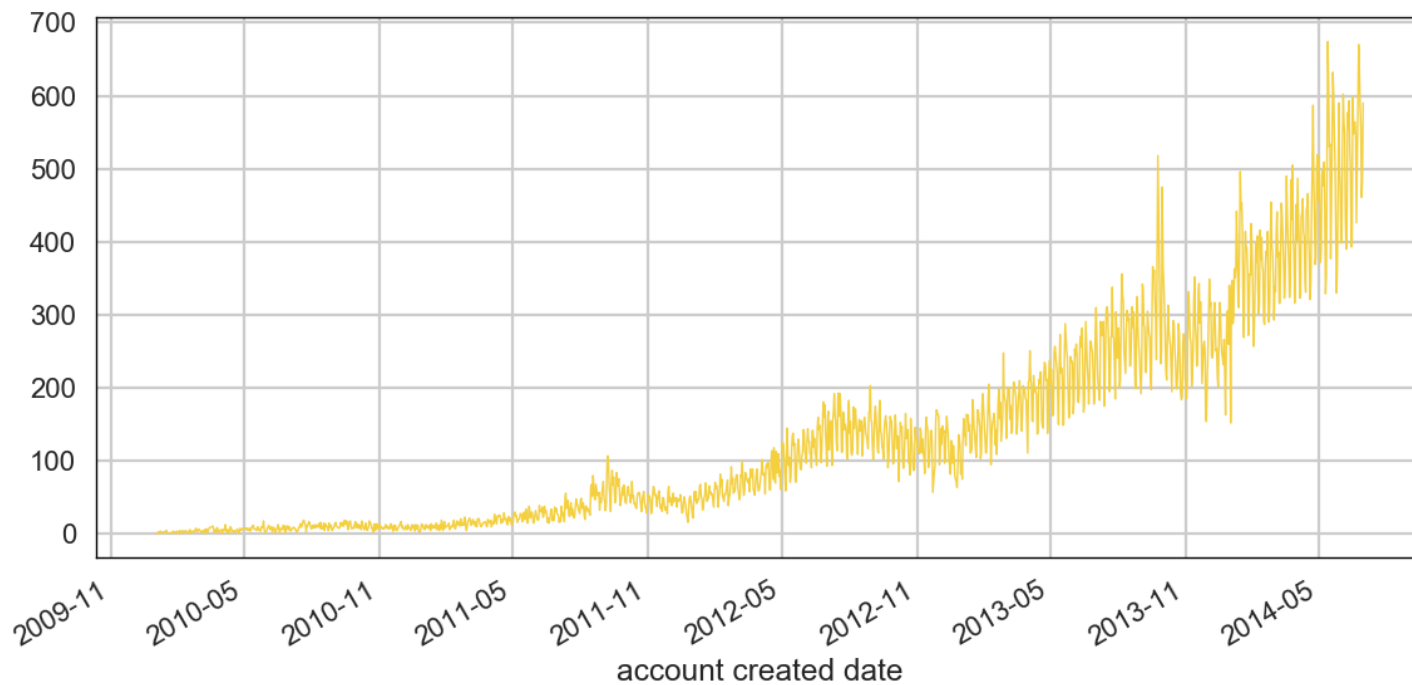
Gender
distribution in
different
destinations

Density of age distribution

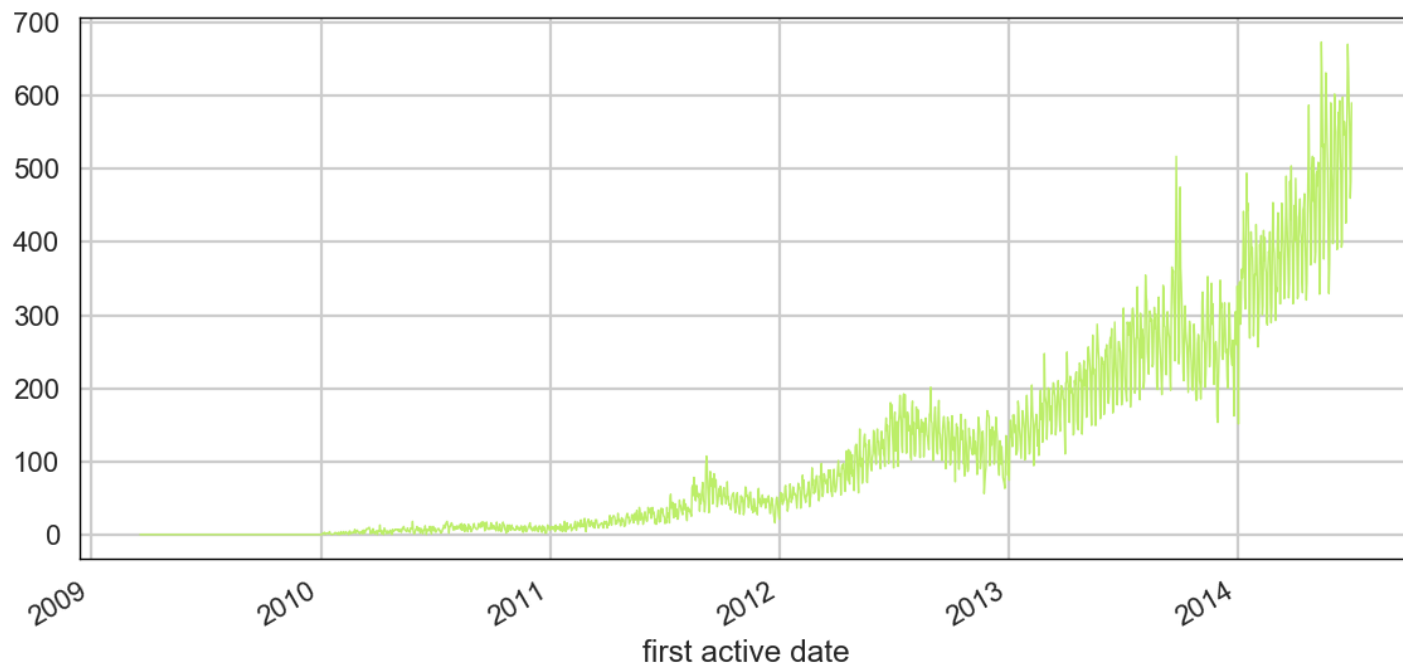


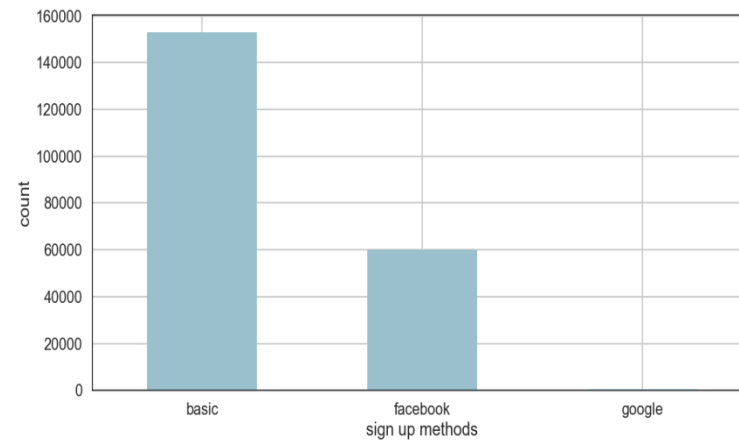
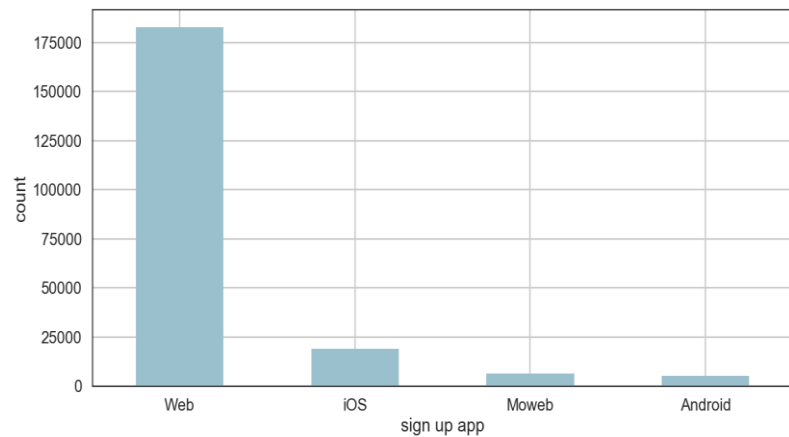
Boxplot of age on different destinations

Distribution of
account create
date

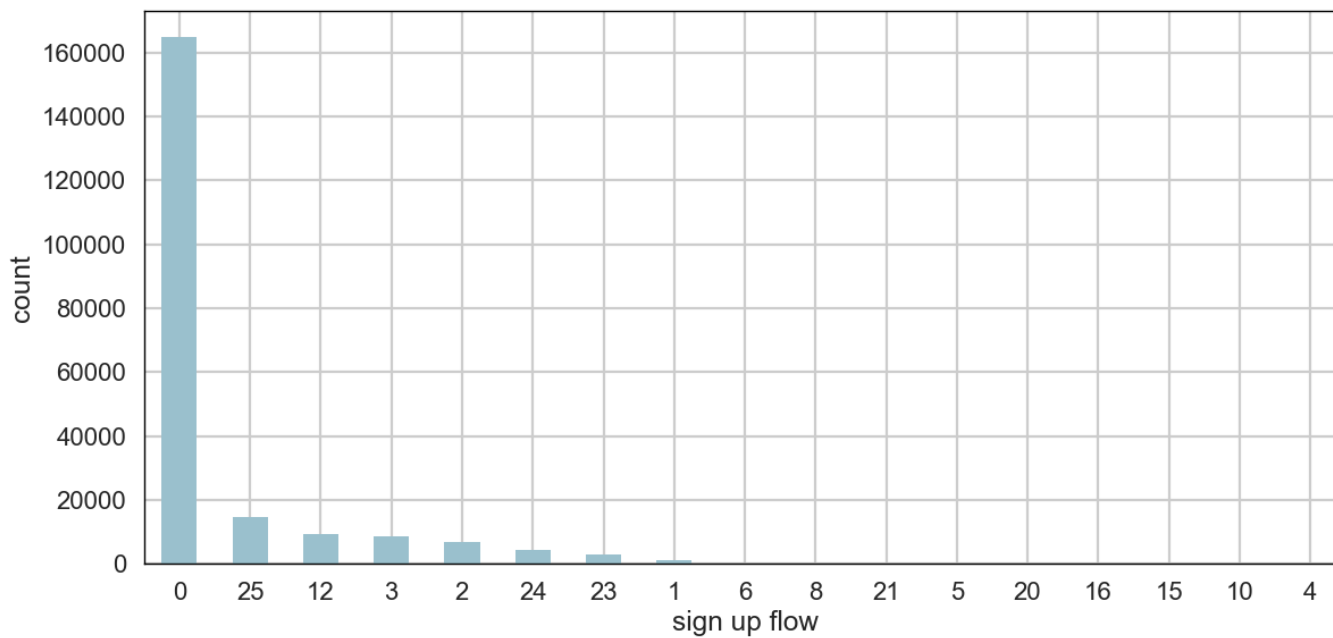


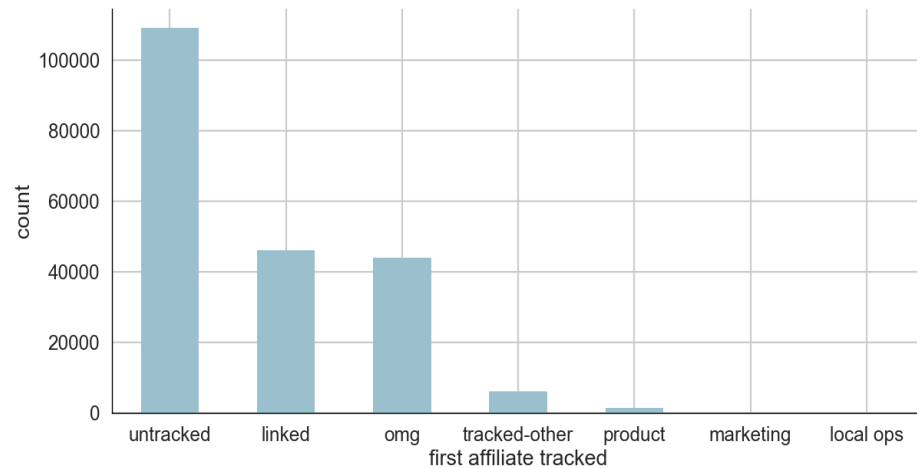
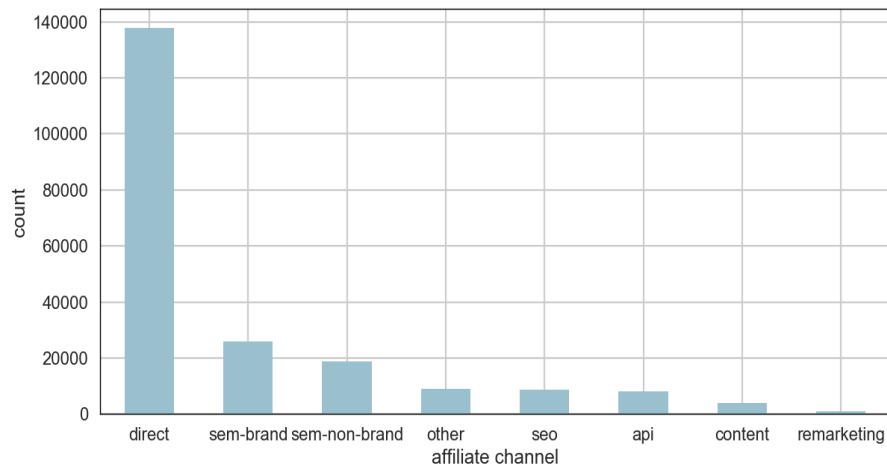
Distribution of
first active date



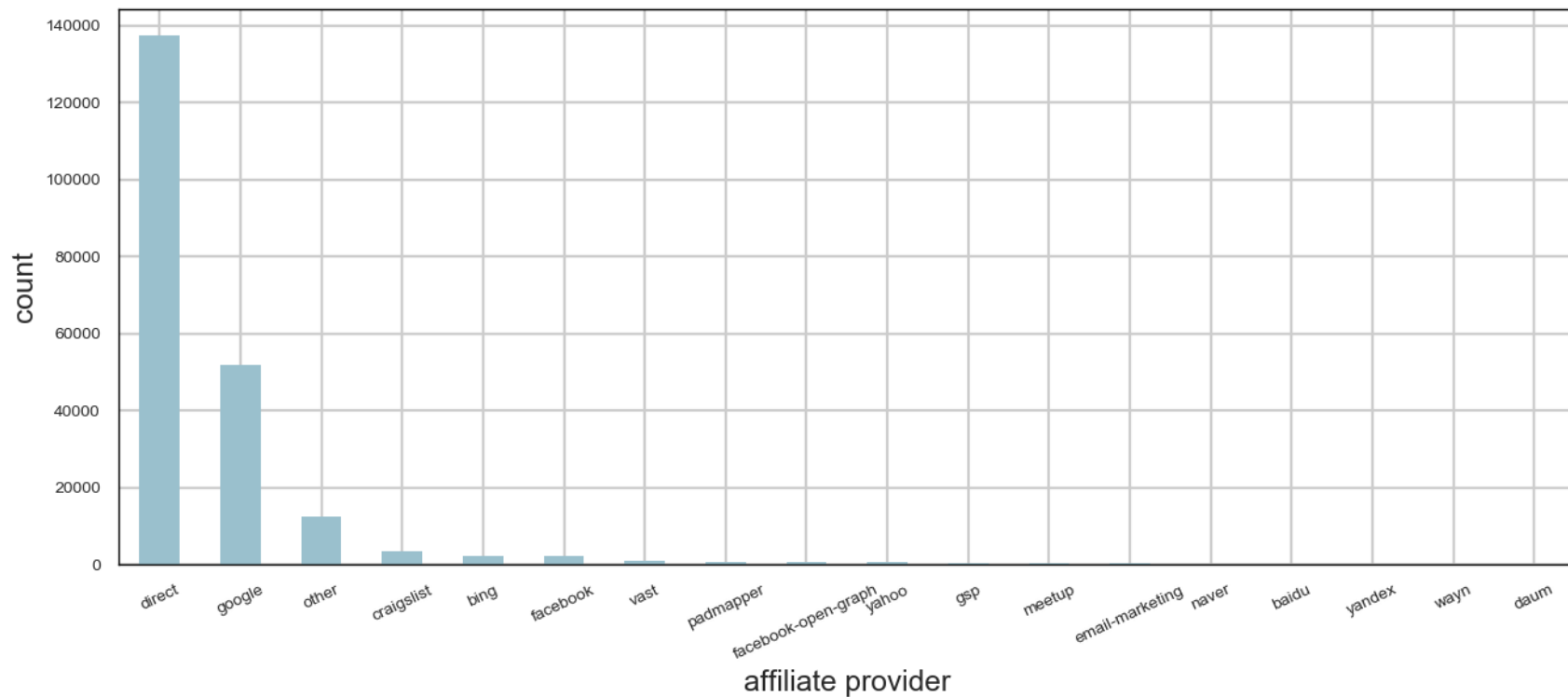


Sign up features

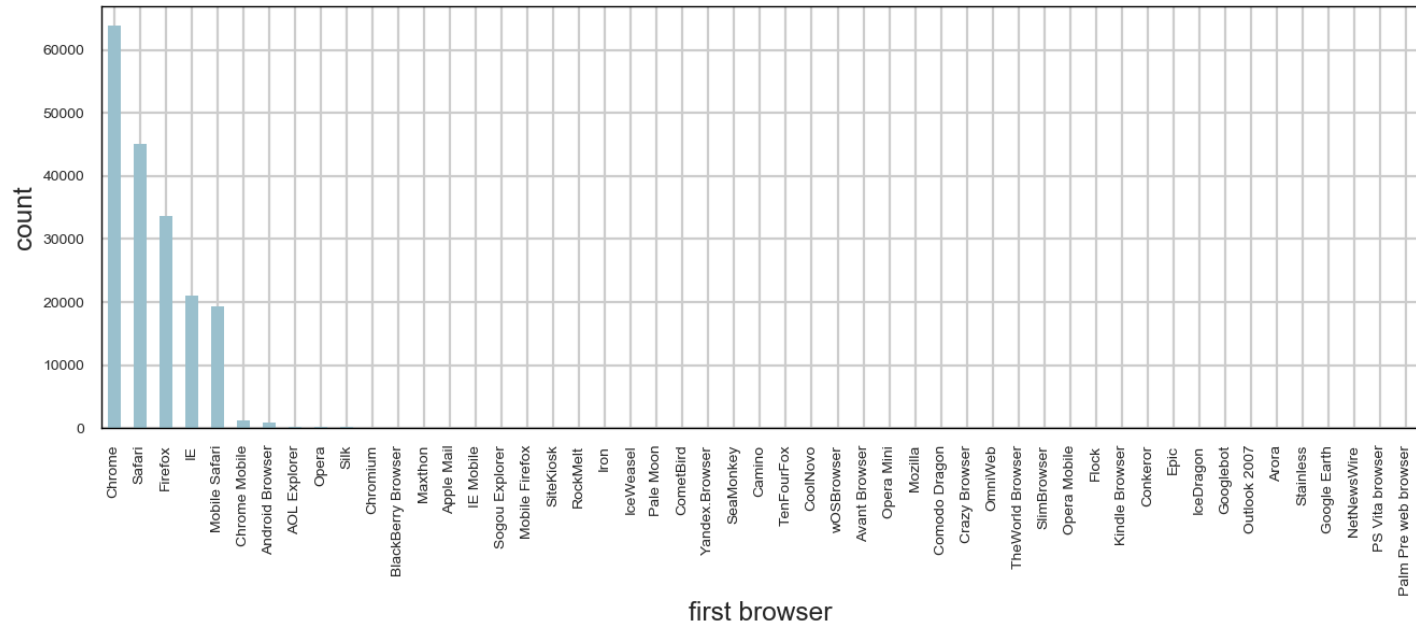
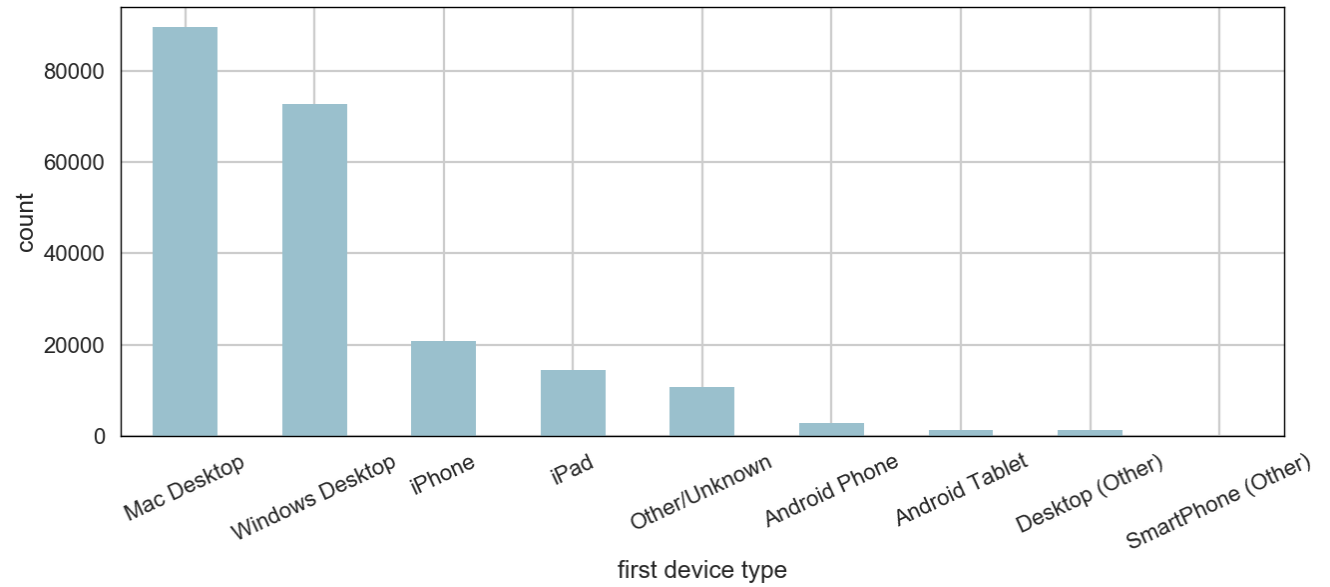




Affiliate features



First device type



First browser

Data preprocessing

- NaN processing :
- The gender, language and first_browser with - unknown- and first_affiliate_tracked with untracked ---- transforme into NaN

Data preprocessing

- NaN processing:

Attribute name	Percentage of missing value
Age	42.412365
Country_destination	22.535538
First_affiliate_tracked	2.208335
First_browser	16.111226
Gender	46.990169

Data preprocessing

- NaN processing:
- There are some age values being greater than 1000. This might be caused by using the born year as age. So we transfer these 'born-year' data as 'age' data.

Data preprocessing

- Feature engineering:
- *Date-related attributes*
- *Age attribute*
- *Other dummy attributes*

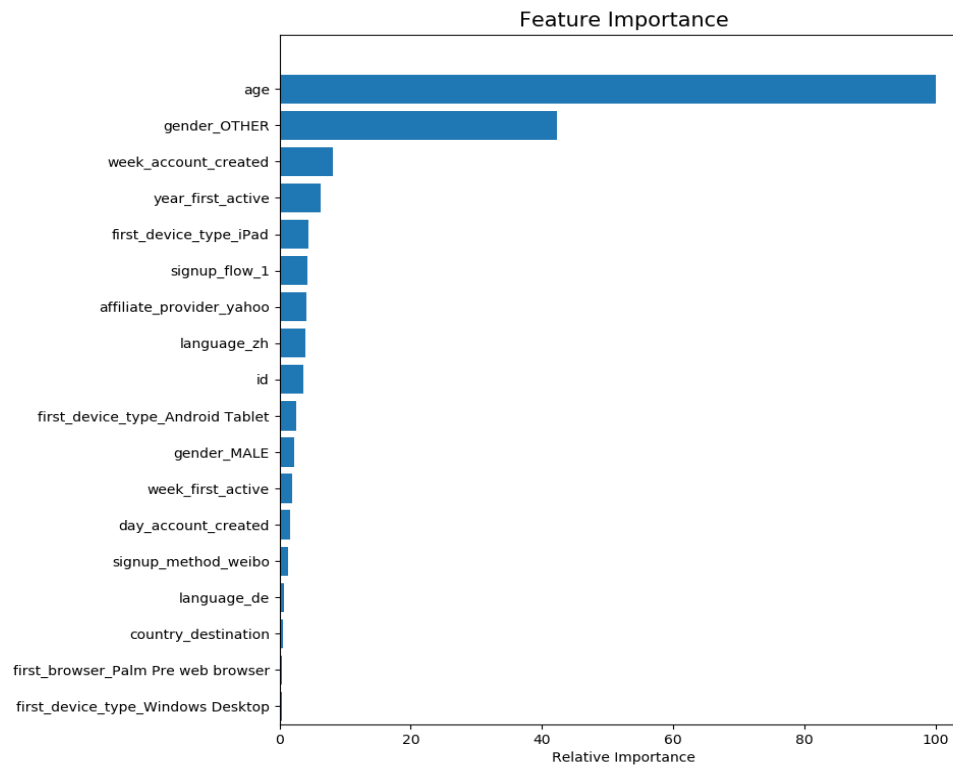
Model Engineering

- Navie Bayes:

True label	NDF	0	0	1	0	0	0	0	167	0	0	0	0
	US	0	0	0	0	0	0	0	431	0	0	0	0
	other	0	0	0	0	0	0	0	297	0	0	0	0
	FR	0	0	0	0	0	0	0	663	0	0	0	0
	CA	0	0	0	0	0	0	0	1530	0	0	0	0
	GB	0	0	0	0	0	0	0	699	0	0	0	0
	ES	0	0	0	0	0	0	0	884	0	0	0	0
	IT	0	0	11	0	0	0	0	37289	0	0	1	0
	PT	0	0	0	0	0	0	0	217	0	0	0	0
	NL	0	0	0	0	0	0	0	70	0	0	0	0
	DE	0	0	21	0	0	0	0	18758	0	0	2	0
	AU	0	0	1	0	0	0	0	2994	0	0	0	0
	Predicted label	NDF	US	other	FR	CA	GB	ES	IT	PT	NL	DE	AU

Model Engineering

- Decision tree----feature importance



Model Engineering

- Decision Tree---- accuracy of entropy and gini index

Results Using Gini Index:

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	168
1	0.00	0.00	0.00	431
2	0.00	0.00	0.00	297
3	0.00	0.00	0.00	663
4	0.00	0.00	0.00	1530
5	0.00	0.00	0.00	699
6	0.00	0.00	0.00	884
7	0.70	0.83	0.76	37301
8	0.00	0.00	0.00	217
9	0.00	0.00	0.00	70
10	0.49	0.51	0.50	18781
11	0.00	0.00	0.00	2995
avg / total	0.55	0.64	0.59	64036

Accuracy : 63.523642950840156

Results Using Entropy:

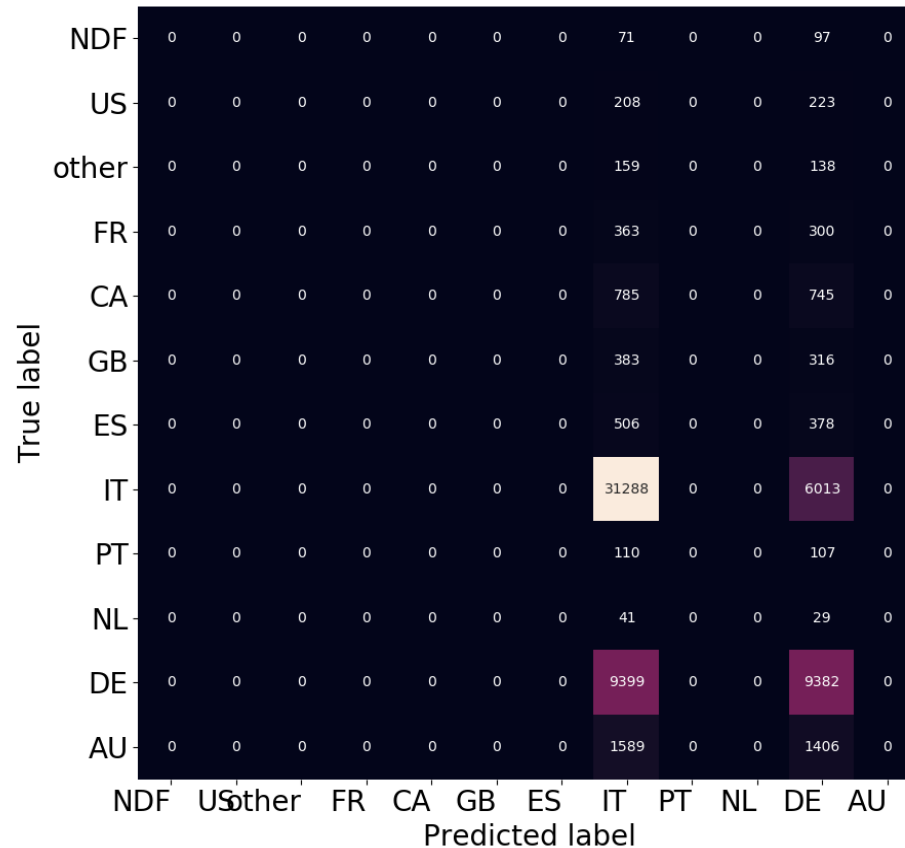
Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	168
1	0.00	0.00	0.00	431
2	0.00	0.00	0.00	297
3	0.00	0.00	0.00	663
4	0.00	0.00	0.00	1530
5	0.00	0.00	0.00	699
6	0.00	0.00	0.00	884
7	0.70	0.84	0.76	37301
8	0.00	0.00	0.00	217
9	0.00	0.00	0.00	70
10	0.49	0.50	0.49	18781
11	0.00	0.00	0.00	2995
avg / total	0.55	0.64	0.59	64036

Accuracy : 63.5111499781373

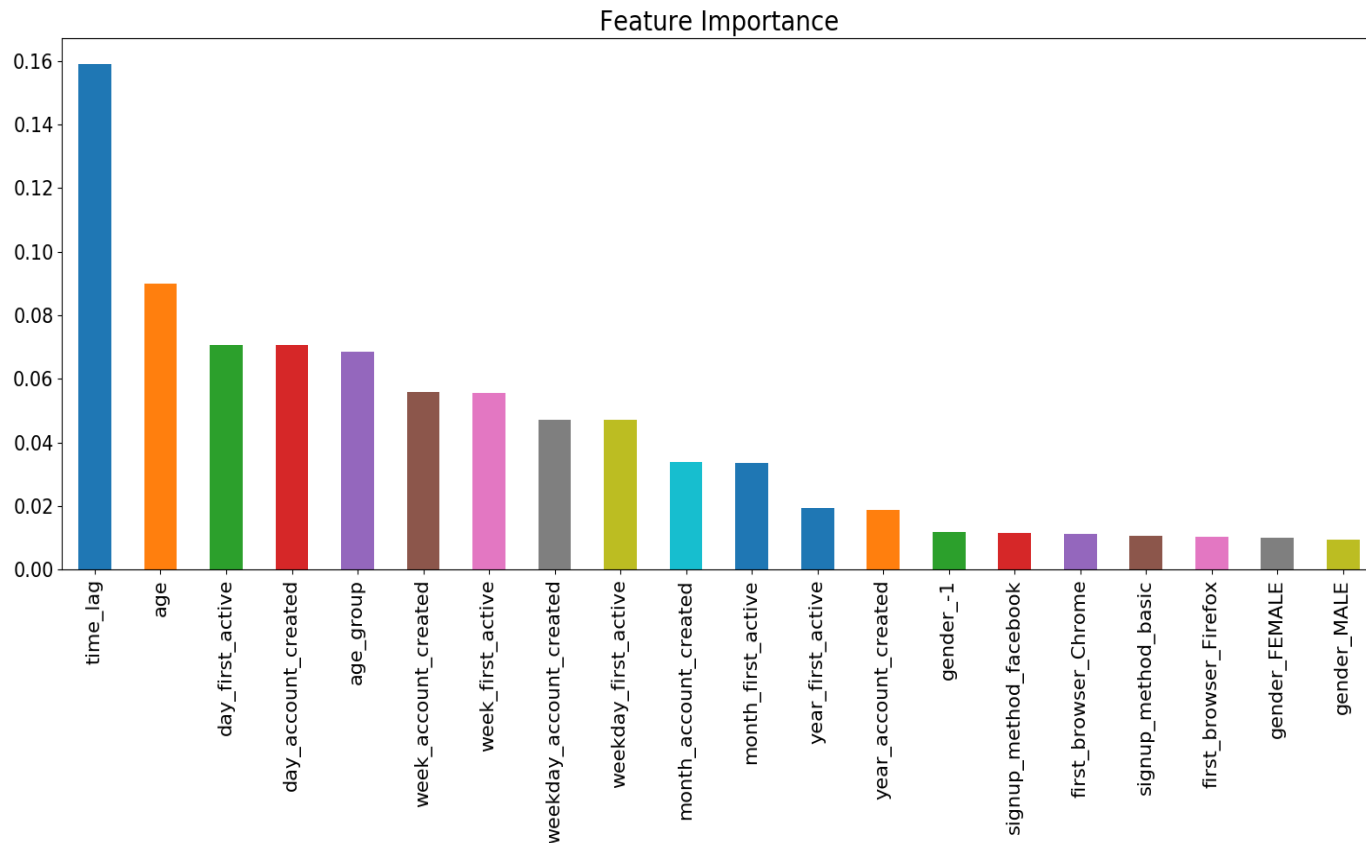
Model Engineering

- Decision Tree---- confusion matrix



Model Engineering

- Random Forest---feature importance



Model Engineering

- Random Forest----accuracy of all features and K features

Results Using All Features:

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	168
1	0.00	0.00	0.00	431
2	0.00	0.00	0.00	297
3	0.02	0.00	0.01	663
4	0.05	0.01	0.01	1530
5	0.03	0.00	0.01	699
6	0.03	0.00	0.01	884
7	0.67	0.81	0.73	37301
8	0.00	0.00	0.00	217
9	0.00	0.00	0.00	70
10	0.46	0.45	0.45	18781
11	0.06	0.01	0.02	2995
avg / total	0.53	0.60	0.56	64036

Accuracy : 60.10837653819726

Results Using K features:

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	168
1	0.00	0.00	0.00	431
2	0.01	0.01	0.01	297
3	0.01	0.01	0.01	663
4	0.02	0.01	0.02	1530
5	0.02	0.01	0.01	699
6	0.02	0.01	0.01	884
7	0.65	0.70	0.68	37301
8	0.00	0.00	0.00	217
9	0.00	0.00	0.00	70
10	0.39	0.42	0.41	18781
11	0.06	0.03	0.04	2995
avg / total	0.50	0.53	0.52	64036

Accuracy : 53.4402523580486

Model Engineering

- SVM----Computationally consuming

Results Using SVM:

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	168
1	0.00	0.00	0.00	431
2	0.00	0.00	0.00	297
3	0.00	0.00	0.00	663
4	0.00	0.00	0.00	1530
5	0.00	0.00	0.00	699
6	0.00	0.00	0.00	884
7	0.68	0.86	0.76	37301
8	0.00	0.00	0.00	217
9	0.00	0.00	0.00	70
10	0.49	0.45	0.47	18781
11	0.00	0.00	0.00	2995
avg / total	0.54	0.63	0.58	64036

accuracy: 63.155100256105946

conclusion

- Navie Bayes
- Decision tree
- Random Forest
- SVM

Thank you!