**Collège de Bois de Boulogne**

**Transformation et manipulation de données**

**TP 1**

**Exploration et transformation de données**

Date de remise: samedi (23H59)

Soit les datasets fournis sur le répertoire du cours. Ceux-ci décrivent les données fournies par AirBnB concernant des locations faites à différents endroits par différents clients (voir le lien : <https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data>)

On nous indique que pour le dataset de test, on veut prédire le pays de destination pour de nouveaux clients (ceux dont la première activité commence après le 7/1/2014.

**Étape 1** : on considère le fichier *train\_users\_2.csv*

Après étude de ce dataset, indiquer les points marquants de votre exploration. Pour chaque observation, indiquer l’opération à effectuer qui serait la plus appropriée.

De ce fait, dans cette phase d’exploration, on considère les points suivants :

* Quelles sont les descripteurs (colonnes) du dataset?
* Combien d'enregistrements (lignes) ont été fournis ?
* Quel est le format des données. Par exemple, dans quel format les dates sont fournies, existe-t-il des valeurs numériques, à quoi ressemblent les différentes valeurs catégorielles ?
* Y a-t-il des valeurs manquantes?
* Est-ce qu’il y’a des dépendances évidentes au niveau des descripteurs?

D’autres observations sur le dataset qui pourraient être pertinentes ?

**Étape 2** : On considère les fichiers train\_users\_2.csv and test\_users.csv

Dans cette étape, on s’intéresse à implémenter les correctifs soulignés dans l’étape 1.

De ce fait, il serait important de considérer les opérations suivantes :

* Conversion de type/format
* Remplacement de valeurs manquantes
* Correction/Suppression de valeurs aberrantes/erronées
* Correction/Standardisation/Normalisation de données

**Livrable** : Votre jupyter notebook devrait contenir le code et la documentation.

**Important** : l’étape de sélection de feature/feature engineering fera l’objet du TP 2.