**Collège de Bois de Boulogne**

**Transformation et manipulation de données**

**TP 1**

**Exploration et transformation de données**

Date de remise: samedi

Équipe de 2 ou individuel

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 :

* Quels 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?

Oui, il existe plusieurs dépendances évidentes entre ces descripteurs, en particulier dans le contexte d’une analyse des utilisateurs (par exemple, une plateforme d'inscription et de réservation comme Airbnb). Voici les principales relations :

### ****1. Dépendances temporelles****

* **date\_account\_created**, **timestamp\_first\_active**, et **date\_first\_booking** sont directement liés :
  + timestamp\_first\_active est souvent antérieur ou égal à date\_account\_created car il enregistre la première interaction d'un utilisateur avant même la création du compte.
  + date\_first\_booking dépend fortement des deux premiers. Un utilisateur doit d'abord créer un compte (date\_account\_created) avant de pouvoir effectuer une réservation (date\_first\_booking).

### ****2. Dépendance entre le genre (****gender****) et l’âge (****age****)****

* Il est possible qu’il y ait une **corrélation entre l’âge et le genre**, notamment en raison de différences démographiques dans l’adoption du service.
* Certains groupes d’âge peuvent être sur-représentés dans certaines catégories de genre (exemple : jeunes adultes plus actifs sur certaines plateformes).

### ****3. Dépendances liées au mode d’inscription****

* **signup\_method**, **signup\_flow**, et **signup\_app** :
  + signup\_method (Google, Facebook, email…) influence signup\_flow (flux de navigation utilisé) et signup\_app (web, mobile…).
  + Un utilisateur s’inscrivant via une application mobile aura souvent un signup\_flow différent de celui qui s’inscrit sur un site web.

### ****4. Dépendance entre**** affiliate\_channel****,**** affiliate\_provider ****et**** first\_affiliate\_tracked

* **affiliate\_channel** (source principale de trafic) et **affiliate\_provider** (partenaire spécifique) sont liés, car certains **fournisseurs sont spécialisés dans certains canaux**.
* **first\_affiliate\_tracked** (si le premier contact avec l’utilisateur via un affilié a été suivi ou non) est aussi influencé par ces deux variables.

### ****5. Dépendance entre**** first\_device\_type ****et**** first\_browser

* Un **type d’appareil** (first\_device\_type) influence fortement le **navigateur** (first\_browser).
  + Ex : Si first\_device\_type = iPhone → first\_browser a de fortes chances d’être Safari.
  + Ex : Si first\_device\_type = Desktop → first\_browser peut être Chrome, Firefox, Edge, etc.

### ****6. Dépendance entre**** language ****et**** country\_destination

* La **langue utilisée (language) influence fortement la destination (country\_destination)**.
  + Un utilisateur qui s’inscrit avec la langue espagnole a plus de chances de réserver en Espagne ou en Amérique latine.
  + Certains pays sont **associés à des langues spécifiques**, bien que des exceptions existent.

## ****📌 Conclusion****

Ces relations peuvent être utilisées pour :

* **Détecter des incohérences** dans les données (ex. : timestamp\_first\_active après date\_account\_created serait suspect).
* **Optimiser les modèles de prédiction** en tenant compte de ces dépendances.
* **Faire du feature engineering** pour créer de nouvelles variables plus pertinentes.

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
* etc…

**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.