# Exploration de Données de GreenDwell pour aider à la prise de Décision

## I - Description du projet et Fichiers de travail

### 1. Contexte du projet

Vous venez d'être embauché en tant que analyste de données junior chez GreenDwell, une entreprise à taille humaine innovante et spécialisée dans les solutions durables pour l'habitat. Elle s'efforce de promouvoir les bâtiments écologiques et économes en énergie en fournissant des solutions intelligentes pour les maisons et les bureaux. Pour se faire, elle a mis à disposition de ses clients une plateforme en ligne pour accéder à un ensemble des services qu'elle propose.

Un client inscrit sur cette plateforme devra renseigner les informations suivantes : nom et prénom et email, et il doit choisir un plan d'abonnement (Éco Basique à 2.99€ est le plan par défaut, Éco Confort à 7.99€ et Éco Prémium à 12.99€).
Ces plans offrent un ensemble de services tels que :

→ **Plan Éco Basique** : accès à une bibliothèque de contenu éducatif (articles, vidéos et webinaires sur la durabilité, l'économie d'énergie et les pratiques écologiques à la maison), analyse énergétique de base et proposition d'un rapport annuel simplifié sur la consommation d'énergie au sein du foyer et des recommandations générales pour l'amélioration, remises sur une gamme limitée de produits éco-responsables vendus par GreenDwell ou ses partenaires (produits d'éclairages LED, systèmes de récupération des eaux de pluie, appareils électroménagers éco-énergétique, panneaux solaires, produits d'isolation, etc), ….

→ **Plan Éco Confort**: consultations personnalisées (par téléphone ou en ligne avec des experts en durabilité pour discuter des améliorations spécifiques de l'habitat, audits énergétique à domicile pour identifier les fuites d'énergie avec recommandations personnalisées, accès prioritaire à des événements (ateliers, séminaires), offres spéciales partenaires sur une large gamme de produit incluant des technologies vertes avancées, etc.`

→ **Plan Éco Prémium**: gestion d'un projet pour la rénovation éco-responsable en mettant en œuvre des solutions incluant la coordination avec des artisans et des fournisseurs spécialisés, suivi énergétique continu grâce à l'installation de dispositifs de suivi de l'énergie pour une surveillance en temps réel de la consommation d'énergie, avec des rapports détaillés et des recommandations personnalisées pour optimiser l'efficacité énergétique, programme de fidélité et récompenses offrant des avantages exclusifs, tels que des mises à niveau gratuites de produits, des services supplémentaires sans coût additionnel, et des invitations VIP à des événements sur le thème de l'écologie, accès en avant-première à des technologies émergentes dans le domaine de l'habitat durable, telles que des systèmes de gestion de l'énergie à domicile intelligents ou des solutions de production d'énergie renouvelable.

Sur la plateforme, on dispose d'un ensemble de données concernants les clients, les services offerts, les sessions de login des clients et l'historique d'accès à la plateforme donnant un aperçu de la manière dont les clients utilisent leur abonnement et interagissent avec les différents services offerts.

## 2. Fichiers de travail

Vous disposez deux fichiers suivants :

- **`clients_greendwell.csv`**
- **`historique_acces_greendwell.xlsx`** (contenant trois feuilles)

Le fichier de clients contient les descripteurs suivants :

→ **`ID client`**, **`Nom Client`**, **`Email`**, **`Date Inscription`**, **`Plan`**, **`Taux Abonnement`** (le coût), **`Remise`** (appliquée ou non à l'abonnement) et **`Date Annulation`**.

Le fichier de l'historique d'accès à la plateforme contient trois feuilles :

→ feuille des services consommés dont les descripteurs sont : **`ID Service`**, **`Nom Service`**, **`Type Service`** (Basique, Confort ou Premium) et **`Popularite`** (mesure de combien de fois le service a été utilisé ou demandé par les clients).

→ feuille sur les sessions de consultation dont les descripteurs sont : **`ID Session`**, **`Date Heure`** et **`Nombre Interactions`** (Combien d'actions le client a effectuées pendant la session, par exemple, pages vues, clics sur des produits, demandes d'informations).

→ feuille sur l'historique de navigation dont les descripteurs sont : **`ID Session`**, **`ID Client`**, **`ID Service`** et **`Ordre Service`** (la séquence dans laquelle les services ont été utilisés pendant une session).

**Remarque :** Ces fichiers sont disponibles sur l'**`Espace Partagé, STID/SD3/S6.`**

### 3. Problématique et mission

L'équipe de direction a remarqué que malgré l'intérêt croissant pour la durabilité, l'entreprise perd de plus en plus des clients. Elle aimerait utiliser la science des données pour comprendre comment réduire le taux de résiliation des clients.
En tant qu'analyste de données chez GreenDwell, vous aurez accès à un ensemble de données sur les interactions des clients avec la plateforme. Votre tâche consiste à collecter les données, à les nettoyer et les explorer pour fournir des aperçus sur les problèmes récents de perte de clients, puis à préparer ces données pour une modélisation future. Votre objectif final est de fournir à l'équipe de direction des insights (observations) basés sur les données qui peuvent aider à optimiser des stratégies de marketing ou autres pour attirer de plus en plus les clients.

## II - Bibliothèques utilisées et étapes de mise en oeuvre

La solution à mettre en œuvre devra être implémentée en utilisant Python et les modules **`Numpy`**, **`Pandas`** et **`Matplotlib`**.

Il faut utiliser GIT pour un échange de versions de code et adopter l'approche incrémentale du développement logiciel.

Voici un descriptif de l'ensemble des étapes à suivre :

### → Étape N°1 : Collecter les données

Lire les deux fichiers **`.csv`** et **`.xlsx`** et les stocker dans des **`DataFrames`**. Chacun représente une entité donnée (clients, sessions, etc).

### → Étape N°2 : Nettoyer les données

Il faut s'inspirer des exercices traitées en TP. Cette étape consiste à :

- Convertir les données des tables aux types appropriés (vers un numérique ou vers une date)
- vérifier s'il y a des données manquantes, des incohérences ou typos, des lignes dupliquées ou des valeurs aberrantes et corriger ces anomalies. Il faut mobiliser les méthodes adéquates de Pandas pour effectuer cette vérification. Par exemple, **`df.info()`** donne un aperçu sur le nombre de valeurs manquantes.

➢ Ajout de nouvelles colonnes si nécessaire telles que : **`Resiliation`** (pour savoir si un client à résilier ou non).

### → Étape N°3 : Explorer vos données

➢ Essayer de mieux comprendre les clients qui ont résilié. Par exemple : depuis combien de temps étaient-ils membres avant d'annuler ? Quel est le pourcentage de clients ayant une remise et ayant résilié leur abonnement par rapport aux clients qui n'ont pas annulé ?

➢ Essayer de mieux comprendre l'historique de navigation des clients. Par exemple, essayer de joindre ensemble les tables d'historique et des services, puis calculer le nombre de sessions de chaque client ainsi que les services les plus utilisés.

Dans cette étape, il faut mettre en œuvre des visualisations et interpréter les résultats.

### → Étape N°3 : Préparer vos données pour la modélisation

➢ Il faut créer un **`DataFrame`** prêt pour la modélisation avec chaque ligne représentant un client et les colonnes suivantes, numériques et non nulles : ID Client, Si un client a annulé ou non, Si un client a reçu une remise ou non, Le nombre de sessions, Pourcentage de l'historique du plan basique, Pourcentage de l'historique du plan confort et Pourcentage de l'historique du plan premium.

➢ Calculer les corrélations et proposer des observations en se basant sur les bonnes variables permettant de prédire la résiliation d'un client.

## III - Modalités du rendu de projet et rapport

Voici les principales modalités à respecter :

**1- Fichiers Python** contenant les **Docstrings et commentaires pertinents** :
Chaque fonction/classe/méthode doit être documentée avec des ***docstrings***. Une chaîne documentaire doit expliquer ce que fait la fonction/classe/méthode, quels sont ses arguments, et ce qu'elle renvoie. Cela aide le correcteur et les autres membres de votre équipe à comprendre et à utiliser vos fonctions. En plus des docstrings, vous devriez ajouter des commentaires pertinents dans le code pour expliquer les parties complexes ou importantes de votre implémentation.
**Les commentaires insérés devront également expliquer votre analyse des résultats obtenus pour chaque action des étapes expliquées ci-dessus.**

**2- Qualité du code** :
Vous devez éviter la duplication de code autant que possible. Si plusieurs fonctions ont besoin de faire la même chose, envisager de créer une fonction utilitaire pour éviter de

répéter le code. Utiliser des noms de variables/classes/méthodes clairs et explicites pour que le code soit facile à comprendre. Éviter les noms de variables obscurs ou trop courts. Surtout adopter une approche orientée objet.

### 3- Dépôt sur Ecampus (jusqu'au 17/03 à 18h)

Une archive compressée : **`NomV_NomW.zip`** (**`V, W, X, Y`** et **`Z`** sont les noms des étudiants) contenant les **fichiers Python y compris docstrings et commentaires**. **Les groupes dont le travail est non soumis pour évaluation auront 0 comme note finale de projet. Chaque retard ou copies de codes (codes identiques) engendrent des points en moins.**

### 4- Rapport (15/12) :

**Chaque groupe devra** rendre un rapport écrit (en format .pdf) contenant les éléments suivants :

→ **Une page de garde** et un **sommaire** communs.

→ **Introduction commune**

→ **Conclusion individuelle** (chaque étudiant devra rédiger sa propre conclusion avec des perspectives d'amélioration du projet).

→ **Développement et mise en œuvre** :

**Une première partie commune** consiste à illustrer l'application de l'approche incrémentale et itérative du développement logiciel, en décrivant comment votre groupe a progressivement construit et amélioré la solution.

Ensuite chaque étudiant devra rédigée sa propre section de développement. Cette dernière devra inclure les éléments suivants pour chaque DataFrame utilisé (clients, sessions, historique d'accès) :

-**Nettoyage de données** : Description des étapes de nettoyage effectuées, telles que la gestion des valeurs manquantes, la correction des erreurs de saisie, la standardisation des formats de données et l'élimination des doublons. Justification des choix effectués lors du nettoyage, en expliquant pourquoi certaines données ont été modifiées ou supprimées.

-**Analyse Exploratrice des données** : Présentation des statistiques descriptives pour comprendre la distribution et les tendances des variables clés, visualisations des données pour illustrer les points importants de l'analyse, avec des graphiques appropriés, identification et discussion des relations potentielles entre les variables, en mettant l'accent sur les facteurs pouvant influencer la fidélité des clients.

-**Conclusions et recommandations** : Pour chaque étape d'analyse, résumez les observations clés, en soulignant leur pertinence pour la problématique mentionnée.

Si vous rencontrez des obstacles lors de l'implémentation de la solution, il est important de les détailler et d'expliquer les mesures prises pour les surmonter.

Afin de simplifier la collaboration en équipe, il est recommandé de partager les différentes versions du code via Git et de joindre à ce rapport des preuves de ces échanges, telles qu'une capture d'écran montrant l'organisation des branches ou d'autres éléments justificatifs.