INF8100 - Concepts et techniques de la fouille et de l'exploitation de données

Travail Pratique 2 préparé par Nairouz Mrabah et Ange Tato

Automne 2023

1 Consignes de remise du travail

Le travail doit être remis au plus tard le **26 Novembre (23:59 EDT) 2023** via *Moodle*. **Important** : Pour chaque jour de retard, vous perdrez 5% de votre note. Après 7 jours vous aurez un 0. Pas de période de grâce une fois le délai écoulé. Pas de période de grâce une fois le délai écoulé. Le travail peut se faire en **équipe de 2 au maximum**. Votre remise doit être un fichier **.zip (-2 pts si ce n'est pas le cas)** qui contient :

- Un rapport (PDF) contenant les réponses aux questions. Il doit y avoir votre nom et votre code permanent.
- Deux fichiers `.ipynb` contenant le code qui vous a permis de répondre aux questions des deux parties 1 et 2. Les réponses aux questions doivent être bien identifiées (numéro). Toutes les réponses doivent être justifiées par un code écrit.
- Deux fichiers `.csv` bruts que vous avez ratisser.

2 Critères d'évaluation

- Réponses aux questions : **40/40**
- La présentation du rapport, et les deux fichiers .ipynb : bonus de **2/40**

Ce TP est noté sur 42 (il est possible d'avoir 41/40 ou 42/40, dans ce cas le bonus ne serait pas comptabilisé dans la note finale) et compte pour 20% de votre note finale.

3 Objectif

Le but de ce travail est de ratisser le web afin de collecter des données, les préparer, les visualiser et appliquer des algorithmes de régression. Il nécessite

un certain apprentissage individuel du langage python et de la librairie beautifulsoup.
Remarque: Le ratissage ne doit faire intervenir aucun API qui permet de faciliter la collecte de données. Utiliser le langage python pour trouver des réponses à toutes les questions ci-dessous (c'est-à-dire, ne faites aucun calcul à la main). Mettre uniquement les réponses (pas de code) dans le rapport PDF. Il est important de numéroter les questions (utiliser le markdown par exemple) dans votre notebook.

4 Partie 1

Vous allez choisir l'un des 2 sites suivants : duproprio.com, publimaison.ca. Ce sont des sites qui offrent un service de vente immobilière sans intermédiaire. Nous allons nous intéresser à l'ensemble des annonces de **vente** disponible sur le site. Vous devez extraire l'ensemble des annonces disponibles sur ce site et mener un ensemble de prétraitement, visualisation et analyses.

4.1 Collecte de données (6pts) :

1. Est-ce que le ratissage des annonces sur le site web que vous avez choisi est permis ? Justifier votre réponse.

2. Vous devez extraire dans un fichier `.csv` à remettre, l'ensemble des annonces (lancer la recherche sans aucun critère). Le nombre doit être le nombre maximum actuel d'annonces publiées sur le site. Une annonce fait référence à un condo/appartement, un terrain à vendre, une maison, bref tout ce qui est à vendre sur le site. Voici les informations brutes à extraire :

   - **Adresse**;
   - Le **prix** demandé en $;
   - **Ville**; **Remarque**: sur chacun des 2 sites web, les villes ne sont pas vraiment des "villes" au vrai sens du terme. Par exemple, on y retrouvera Anjou ou encore Mont-royal comme villes dans la région Montréal/ l'île. Référez vous à la barre de recherche du site en question pour plus de détails.
   - **Région**; **Remarque**: de même que pour les villes, référez vous à la barre de recherche sur le site pour voir la liste des régions. Exemple: Laurentides, Laval, Montréal/l'île;
   - Le nombre de **Chambres** dans la maison;
   - Le nombre de **salles de bain**;
   - Le nombre de **salles d'eau**;
   - Le nombre d'**étages**;
   - L'**aire habitable** en pi^2;

- La **taille du terrain** en pi^2;
- Le montant annuel des **taxes municipales**;
- Le montant annuel des **taxes scolaires**;
- Le montant annuel de l'**électricité**;
- Le montant annuel des **assurances**.

```
#Les colonnes de votre fichier csv.
columns = ["Adresse", "Prix","Ville","Région","Chambres","Salles de bain", "Salles d'eau", "Étages", "Aire habitable",
"Taille terrain","Taxes municipales","Taxes scolaires","Électricité","Assurances"]
```

4.2 Nettoyage et exploration des données (4pts) :

1. Combien ya-t-il de valeurs manquantes dans chaque colonne de votre jeu de données?

2. Selon vous, quel est la cause de ces valeurs manquantes ? Est-ce que parmi les colonnes qui ont des valeurs manquantes, on pourrait utiliser l'une des techniques de remplacement de valeurs manquantes vues en cours ? Si oui dites pour les colonnes concernées, lesquelles des techniques fonctionneraient bien.

3. Combien ya til de régions différentes ? et de villes différentes ?

4. Quel est le type (inféré par pandas) de données de chaque colonne ?

5. Nettoyer vos données : correction d'erreurs, traitement de valeurs manquantes s'il ya lieu, correction du type des données.

6. Quel est le prix moyen des maisons (au moins 1 chambre et 1 salle de bain) sur l'île de Montréal ? À Laval ? Dans les laurentides ?

7. Dans quelle ville de Montréal/l'Île les maisons (au moins 1 chambre et 1 salle de bain) coûtent le moins chers ?

8. Pour chaque région, afficher le prix de l'item (annonce) le plus élevé et la ville où l'item se situe. Ici on ne fait pas de différence si c'est un condo/appartement, maison, terrain vide, etc. À quel région/ville revient la palme d'or de l'item le plus cher ? Donner toutes les caractéristiques (valeurs de toutes les colonnes) de cet item.

4.3 Visualisation et analyse des données (6pts) :

1. Présenter visuellement (à l'aide d'un graphique) la matrice de corrélation entre les colonnes numériques. Y a-t-il des corrélations de plus de 0.7 ? quelles sont elles ?

2. Présenter visuellement la proportion numérique de chaque région en matière de nombre d'annonces, par rapport à l'ensemble des annonces. Quelle région occupe la plus petite proportion ?

3. À l'aide d'un graphique différent de celui de la question précédente, comparer le nombre d'annonces de vente pour chaque région. Quelle région possède le plus d'annonces de vente ?

4. À l'aide d'un graphique, comparer le prix moyen des annonces pour chaque région. Quelle région possède le prix moyen le plus élevé ?

5. Pour ce point, on se limite aux annonces ayant au moins 1 chambre et 1 salle de bain. À l'aide d'un graphique, comparer le prix moyen de ces annonces pour chaque région. Quelle région possède le prix moyen le plus élevé pour les annonces avec au moins 1 chambre et 1 salle de bain?

6. À l'aide d'un graphique, analyser la relation entre le prix des annonces et le nombre de chambres. Y a-t-il un lien quelconque ? Est-ce que la région y joue un rôle dans cette relation? Peut-on apercevoir des valeurs aberrantes ? Si oui identifiez-les : donnez toutes les valeurs des colonnes de ces valeurs aberrantes.

7. À l'aide d'un graphique, analyser la relation entre la valeur des taxes municipales annuelles des annonces et la taille du terrain. Y a-t-il un lien quelconque ? Est-ce que la région y joue un rôle dans cette relation? Peut-on apercevoir des valeurs aberrantes ? Si oui identifiez-les : donnez toutes les valeurs des colonnes de ces valeurs aberrantes.

8. À l'aide d'un graphique, analyser la relation entre la valeur des taxes municipales annuelles des annonces et le prix. Il y a-t-il un lien quelconque ? Est-ce que la région y joue un rôle dans cette relation? Peut-on apercevoir des valeurs aberrantes ? Si oui identifiez-les : donnez toutes les valeurs des colonnes de ces valeurs aberrantes.

9. On s'intéresse pour cette question aux annonces qui ont un prix affiché de moins de 1 million de $, pour toutes les régions. Dessiner dans un même graphique un boxplot représentant la répartition de prix par région. Analyser de manière détaillée le graphique obtenu.

10. On s'intéresse pour cette question aux maisons de 2 chambres au moins et une salle de bain au moins et qui coûte moins de 1 million de $, pour toutes les régions. Dessiner dans un même graphique un boxplot représentant la répartition de prix par régions. Analyser de manière détaillée le graphique obtenu. Est-ce qu'il y a des différences entre ce graphique et celui de la question précédente ? Si oui donner en 4.

11. En un seul graphique, présenter une analyse bivariée de toutes les colonnes numériques de votre jeu de données. Analyser en détail le graphique obtenu.

4.4 Algorithmes de régression (6pts) :

Dans cette partie, on gardera toujours 85% des données pour l'entraînement et le reste pour les tests. **Remarque**: Choisissez la bonne transformation pour vos données et justifiez vos choix!

1. Dans la matrice de corrélation présentée ci-dessus, identifier 2 variables différentes qui ont le plus haut coefficient de corrélation. Concevez un modèle de régression linéaire dont l'une des valeurs est à prédire et l'autre est la valeur d'entrée. Le modèle de régression construit n'est autre qu'une droite. Vous devez représenter cette droite dans un graphique, ainsi que les points de données qui représentent les 2 variables. Est-ce que la droite telle que présentée sur votre graphique fait une bonne approximation de vos points/données? Vérifier votre réponse avec les données de test.

2. Dans cette question, on s'intéresse à prédire si le prix d'une annonce sera supérieur ou inférieur à 350000$ en fonction de la région, du nombre de chambres, le nombre de salles de bain, le nombre de salles d'eau, le nombre d'étages, la superficie de l'aire habitable, la taille du terrain, les taxes municipales et les taxes scolaires. Concevez un modèle de régression qui permet de faire cette prédiction et évaluer votre modèle.

3. Dans cette question, on s'intéresse à prédire le prix d'une annonce en fonction de la région, du nombre de chambres, le nombre de salles de bain, le nombre de salles d'eau, le nombre d'étages, la superficie de l'aire habitable, la taille du terrain, les taxes municipales et les taxes scolaires. Concevez un modèle de régression qui permet de faire cette prédiction et évaluer votre modèle.

4. Le couple *Formidable* aimerait vendre 2 de ses propriétés. En vous servant de votre modèle construit ci-dessus, à combien est estimé le prix de vente de chacune des deux propriétés ? Voici les caractéristiques :

   - **Propriété 1**: région: Québec Rive-Nord, nombre de chambres: 3, nombre de salles de bain: 2, nombre de salles d'eau: 1, nombre d'étages: 2, superficie de l'aire habitable: 1700.2 pi^2, taille du terrain: 5060 pi^2, taxes municipales: 4272,39$, taxes scolaires: 411,06$, électricité: 3 584,00 $, assurances 110,38 $.
   - **Propriété 2**: ville: Ferme-Neuve, région: Laurentides, taille du terrain 8021.06 pi^2, taxes municipales: 2 324,75 $, taxes scolaires: 65,59$

5. Sans toutefois implémenter, pensez-vous que rajouter la ville dans vos 2 derniers modèles de régression conçue améliorerait la prédiction ? Justifiez votre réponse (un graphique ou un calcul).

5 Partie 2

Dans cette partie, on s'intéresse au siteweb: imdb.com. Internet Movie Database (littéralement, Base de données cinématographiques d'Internet), abrégé en IMDb, est une base de données en ligne sur le cinéma mondial, sur la télévision, et plus secondairement les jeux vidéo. IMDb restitue un grand nombre d'informations concernant les films, les acteurs, les réalisateurs, les scénaristes et toutes personnes et entreprises intervenant dans l'élaboration d'un film, d'un téléfilm, d'une série télévisée ou d'un jeu vidéo. L'accès aux informations publiques est gratuit.

5.1 Collecte de données (6pts) :

Vous devez extraire dans un fichier `.csv` à remettre, l'ensemble des films qui ont les deux caractèristiques suivantes (Title Type="Feature Film" et Release Date="2018-01-01, 2018-12-31"). Utilisez la barre de recherche pour trouver ces films ou utilisez directement le lien suivant: lien. **Remarque**: pour éviter tout blocage possible, vous devez vous servir de la bibliothèque fake-useragent pour formuler les entêtes des requêtes HTML. De plus, il faut prévoir un temps d'attente avec un minimum de 30 seconds entre deux requêtes consécutives.

```
# Création des entêtes
!pip install fake-useragent
from fake_useragent import UserAgent
user_agent = {'User-agent': UserAgent().random}
page = requests.get(url, headers = user_agent).text

# Attendre au moins 30 seconds
sleep_time = 28 + 20 * random.random()
time.sleep(sleep_time)
```

Voici les informations brutes à extraire :

- **id_film:** l'identifiant du film qui est inscrit dans le lien. Vous pouvez utiliser la fonction `extraire_id_film(lien)` (définie en bas) pour extraire l'identifiant du film;
- **titre_film:** le titre du film;
- **durée:** la durée en minutes du film;
- **genres_liste:** une liste contenant les genres possibles du film;
- **date_sortie:** date de sortie du film;
- **utilisateur_note:** la note moyenne du film donnée par les utilisateurs IMDb;
- **nbre_utilisateur_note:** le nombre d'évaluations du film sur IMDb;

```
def extraire_id_film(lien):
    regex = re.compile("^/titre/tt(\d+)/.+$")
    return re.findall(regex, link)


#Les colonnes de votre fichier csv.
columns=['id_film', 'titre_film', 'durée', 'genres_liste', 'date_sortie', 'utilisateur_note', 'nbre_utilisateur_note']
```

5.2 Exploitation des données (12pts):

1. Nettoyer et coder vos données : correction d'erreurs, traitement de valeurs manquantes s'il y a lieu, éliminations des duplications, éliminations des lignes avec des valeurs abérantes, et correction du type des données (codage si c'est nécessaire).
   Remarques non ordonnées:

   - Supprimer les films dont la durée n'est pas enregistrée.
   - Convertir la durée de chaque film en minutes (entier).
   - Supprimer les films dont la durée est égale à 0.
   - Supprimer les films dont la durée est très élevée.
   - Convertir le nombre d'évaluations de chaque film **nbre_utilisateur_note** à une valeur entière.
   - Supprimer les films qui ont un nombre d'évaluations très élevée.
   - Créer une colonne pour chaque genre. Il ne faut laisser que les 5 genres les plus cités et regrouper le reste dans une colonne **autres_genres**.
   - Convertir **date_sortie** au format `datetime`.
   - Supprimer les films qui n'ont pas de date de sortie enregistrée.

2. Créer 2 nouvelles colonnes **durée_minutes_log**, **nbre_utilisateur_note_log**.

   - Appliquer la fonction logarithmique sur la colonne **durée** pour avoir la nouvelle colonne **durée_minutes_log**.
   - Appliquer la fonction logarithmique sur la colonne **nbre_utilisateur_note** pour avoir la nouvelle colonne **nbre_utilisateur_note_log**.

3. Réaliser une analyse univariée complète avec les visualisations adéquates et interpréter les résultats.

4. Réaliser une analyse bivariée complète avec les visualisations adéquates et interpréter les résultats.

5. Dans cette question, on s'intéresse à prédire la note d'un film **utilisateur_note** en fonction de 6 colonnes **nbre_utilisateur_note_log**, **durée_minutes_log**, **drame**, **action**, **thriller**, et **horreur**. Concevez un modèle de régression linéaire qui permet de faire cette prédiction, vérifier les 4 conditions nécessaires pour appliquer la régression linéaire et évaluer votre modèle.

6. Refaire la question précedente après avoir effectuer une normalisation adéquate. Comparer les résutlats.

   Remarques: Pour les deux dernières questions, on garde 80% des données pour l'entraînement et le reste pour les tests.