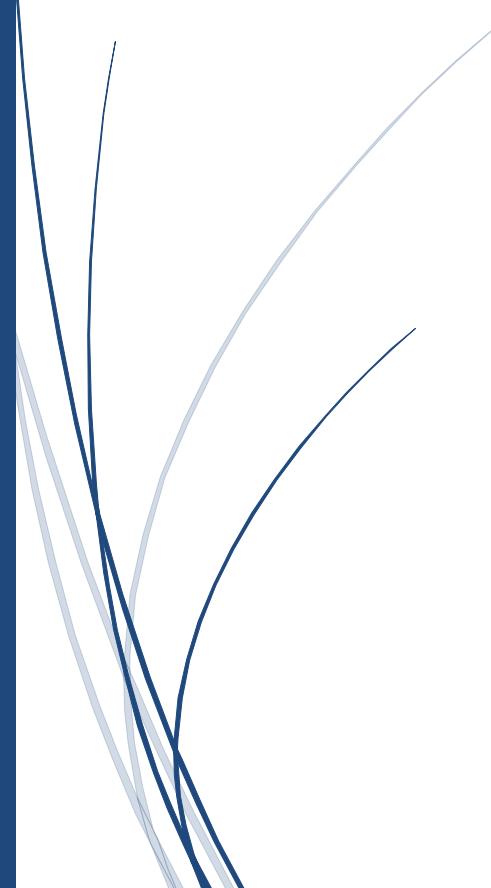


11/28/2022

Travail pratique # 2

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

Chargé de cours : **Nairouz Mrabah**

A decorative graphic consisting of several thin, curved lines in shades of blue and grey, originating from the bottom left corner and extending upwards and to the right.

Thomas Thack Luangxay
Bouaoune Youghourta
Hammami Amine

TABLE DES MATIÈRES

1.	PARTIE 1 : 4.1 COLLECTE DE DONNÉES (6PTS).....	1
2.	PARTIE 1 : 4.2 NETTOYAGE ET EXPLORATION DES DONNÉES (4PTS) :	3
3.	PARTIE 1 : 4.3 VISUALISATION ET ANALYSE DES DONNÉES (6PTS) :	6
4.	PARTIE 1 : 4.4 ALGORITHMES DE RÉGRESSION (6PTS) :	11
5.	PARTIE 2 : 5.1 COLLECTE DE DONNÉES (6PTS) :	16
6.	PARTIE 2 : 5.2 EXPLOITATION DES DONNÉES (12PTS) :	17

1. PARTIE 1 : 4.1 COLLECTE DE DONNÉES (6PTS)

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

Oui le site duproprio.com est permis de faire le ratissage, parce qu'il n'y a aucune vérification de différencier de manière automatisée un utilisateur humain d'un ordinateur. Le site est accessible pour tout le monde. C'est un site publique.

1.2 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 pi²;
- La taille du terrain en pi²;
- Le montant annuel des taxes municipales ;
- Le montant annuel des taxes scolaires ;
- Le montant annuel de l'électricité ;
- Le montant annuel des assurances.

Le fichier CSV doit contenir les colonnes suivantes:

1) Adresse: Adresse;

2) Prix: Le prix demandé en \$;

3) Ville: 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.

4) Région: 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;

- 5) Chambres: Le nombre de Chambres dans la maison;
- 6) Salles de bain: Le nombre de salles de bain;
- 7) Salles d'eau: Le nombre de salles d'eau;
- 8) étages: Le nombre d'étages;
- 9) Aire habitable: L'aire habitable en pi2;
- 10) Taille terrain: La taille du terrain en pi2
- 11) Taxes municipales: Le montant annuel des taxes municipales;
- 12) Taxes scolaires: Le montant annuel des taxes scolaires;
- 13) électricité: Le montant annuel de l'électricité;
- 14) Assurances: Le montant annuel des assurances.

```

Numéro de la pages traité est: 629
Numéro de la pages traité est: 630
Nombre de total de pages traité est: 631

```

```

    Étages Aire habitable Taille terrain Taxes municipales Taxes scolaires \
0      NaN      NaN      3,700,000      NaN      NaN
1      NaN      NaN      1,429,232.03      NaN      NaN
2      NaN      NaN      34,000      400,00$      NaN
3      NaN      NaN      3,280,068      NaN      NaN
4      NaN      NaN      150500      NaN      NaN
...      ...      ...      ...      ...      ...
6884    NaN      NaN      74,000      NaN      NaN
6885    NaN      NaN      23,000      NaN      NaN
6886    NaN      NaN      40,000      NaN      NaN
6887    1.0      700      10,375      1194,86$      138,60$
6888    NaN      NaN      10,500      NaN      NaN

    Électricité Assurances
0      NaN      NaN
1      NaN      NaN
2      NaN      NaN
3      NaN      NaN
4      NaN      NaN
...      ...      ...
6884    NaN      NaN
6885    NaN      NaN
6886    NaN      NaN
6887    NaN      NaN
6888    NaN      NaN
[6889 rows] x 14 columns]>

```

- En total, il y a **6889** annonces.

2. PARTIE 1 : 4.2 NETTOYAGE ET EXPLORATION DES DONNÉES (4PTS) :

2.1 Combien y a-t-il de valeurs manquantes dans chaque colonne de votre jeu de données?

Il y a beaucoup de valeurs manquantes dans chaque colonne. Voir image ci-joint

```
Adresse      0
Prix         0
Ville        0
Région       0
Chambres     1552
Salles de bain 1257
Salles d'eau  4658
Étages       1879
Aire habitable 2263
Taille terrain 1228
Taxes municipales 1815
Taxes scolaires 2127
Électricité   3536
Assurances    4582
dtype: int64

Nombre total de valeurs manquantes: 24897
```

2.2 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.

Il se peut que le client ait oublié de fournir les informations nécessaires ou le client a volontairement choisi de ne pas renseigner un champ. Il se peut aussi que le vendeur a oublié de saisir les valeurs fournies par client. Oui pour certaine colonne, on peut remplacer les valeurs manquantes par une valeur fixe. Si c'est une valeur numérique, on peut remplacer la valeur manquante par une valeur numérique (zéro), dans le cas d'une chaîne de caractère, on peut remplacer la valeur par une chaîne vide.

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

```
Adresse      object
Prix         object
Ville        object
Région       object
Chambres     float64
Salles de bain float64
Salles d'eau  float64
Étages       float64
Aire habitable object
Taille terrain object
Taxes municipales object
Taxes scolaires object
Électricité   object
Assurances    object
dtype: object
```

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

1) Valeurs manquantes

[illegible]

Il faudrait remplacer tous les nombres -1 par un nombre fixe 0

La taille de terrain devrait être un nombre entier en pied carré.

Aire habitable	Taille terrain	Taxes municipales	Taxes scolaires
-1	321.52x127.95	-1	32,000
904.17	8,600	1512.77\$	-
-1	-1	-1	-
-1	48,000	-1	-
-1	93,000	-1	-

Maison	Prixurdemande	Côte-des-Neiges / Notre-Dame-de-... Boucherville	Montréal / Rive-Sud...
ico - Mo...			
s Châteaux	0.25\$	Lanier	Laurentide
amoureux	0.25\$	St-Julienne	Launaudière
4ormandie	0.25\$	Montmagny	Chaudière-Appalaches
in Keenan	0.48\$	Melbourne	Estrie
a Traverser	0.64\$	Lac-Aux-Sabiez	Mauricie
ée Tessier	0.75\$	Lac-Des-Ecorces	Laurentides
dor Condor	0.85\$	St-Casiste	Launaudière
J Domaine		Lamarche	Saguenay-La-Saint...
Belie-M	1\$	St-Jean-De-Matha	Launaudière
n Belterive		Carignan	Montréal (Rive-Sud-...)
i Tourbière	1\$	Ste-Catherine-de-la-JC	Québec Rive-Nord

Ville	Région	Chambres	Salles de bain
Drummondville (St-Nicéphore)	Centre-du-Québec	2	1
St-Nicolas	Québec Rive-Sud (Lévis)	-1	-1
Mille-Isles	Laurentides	-1	-1
Saint-Laurent	Montréal / Île	7	3
Thetford Mines	Chaudière-Appalaches	-1	-1
Morin-Heights	Laurentides	-1	-1
Morin-Heights	Laurentides	-1	-1
Morin-Heights	Laurentides	-1	-1
Morin-Heights	Laurentides	-1	-1

Tous les montants de prix seront un point au lieu d'une virgule.

Taxes municipales	Taxes scolaires	Électricité	Assurances
4600.00\$	359.00\$	1100.00\$	2800.00\$
-1	-1	-1	-1
-1	-1	-1	-1
6216.00\$	462.00\$	-1	-1
8261.26\$	429.00\$	4256.00\$	4101.67\$
6661.84\$	810.60\$	-1	2409.00\$

Page 4 de 27

L'île-Bizard / Sainte-Genevi...
L'île-Bizard / Sainte-Genevi...
Rimouski (Ste-Odile-Sur-Rimouski)

5) Solution appliquée pour les nettoyages de données sont :

- **Remplace la valeur nulle par 0 pour des colonnes suivantes** : Chambres, salles de bain, salle d'eau, Étages, Air habitable, Taille terrain, taxes municipales, taxes scolaires, électricité et assurances.
- Supprimer les lignes où il n'y a pas de valeur pour la colonne "Aire habitable".
- Remplacer le caractère html ''' par un apostrophe.
- Supprimer les maisons coûtées 1\$
- Supprimer les maisons qui n'ont pas de taxes municipale et taxes scolaires.
- Supprimer les maisons qui n'ont pas de la taille du terrain
- Supprimer les maisons qui n'ont pas de la chambre et les salles d'eau.
- Corriger les types des données.
- Chambre en nombre entier
- Salles de bain en nombre entier
- Salle d'eau en nombre entier
- Étages en nombre entier
- Air habitable en nombre entier
- Taille terrain en nombre entier
- Taxe municipales et taxe scolaires, prix, électricité, assurance seront en nombre réel (\$)

```

Adresse          object
Prix             float64
Ville            object
Région           object
Chambres         int64
Salles de bain   int64
Salles d'eau     int64
Étages          int64
Aire habitable   int64
Taille terrain   int64
Taxes municipales float64
Taxes scolaires float64
Électricité      float64
Assurances       float64
dtype: object

Nombre d'enregistrement avant de nettoyage est: 6889

Nombre d'enregistrement après de nettoyage est: 2987

Enregistrer les données propres dans un fichier CSV 'duproprioFinale.csv' pour faire la validation.

```

Maintenant, les données sont propres et conformes.

2.5 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 ?

```

Le prix moyen des maisons sur l'île de Montréal est 564719.96
Le prix moyen des maisons à Laval est 480896.6666666667
Le prix moyen des maisons dans les Laurentides est 578675.0

```

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

	Ville	Prix	Chambres
2174	Pointe-Aux-Trembles / Montréal-Est	339000.0	2

2.7 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. A 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.

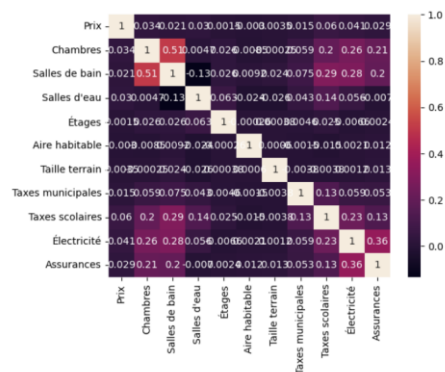
```

Région: Outaouais
Adresse: 60, rue de l'Allée
Prix: 999999.0
Ville: Val-Des-Monts
Chambres: 5
Salles de bain: 1
Salles d'eau: 1
Étages: 2
Aire habitable: 2000
Taille terrain: 6534000
Taxes municipales: 266566.0
Taxes scolaires: 56000.0
Électricité: 320000.0
Assurances: 0.0
  
```

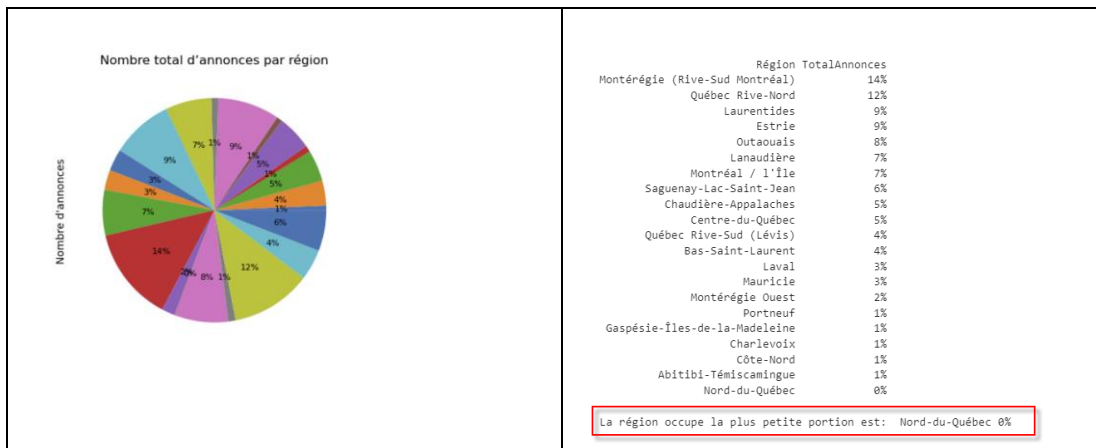
3. PARTIE 1 : 4.3 VISUALISATION ET ANALYSE DES DONNÉES (6PTS) :

3.1 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 ?

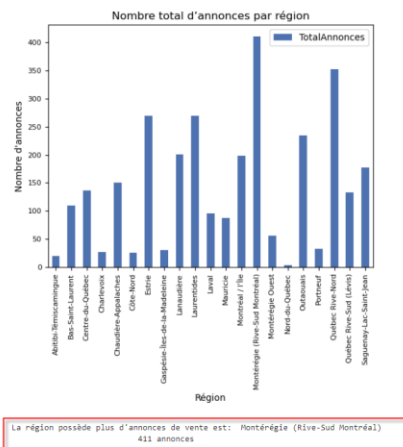
- Selon le graphique, il n'existe aucune corrélation de plus de 0.7. Les corrélations sont tout presque zéro et la valeur maximale est 0.51 (corrélation maximale est située entre salle de bain & chambres)



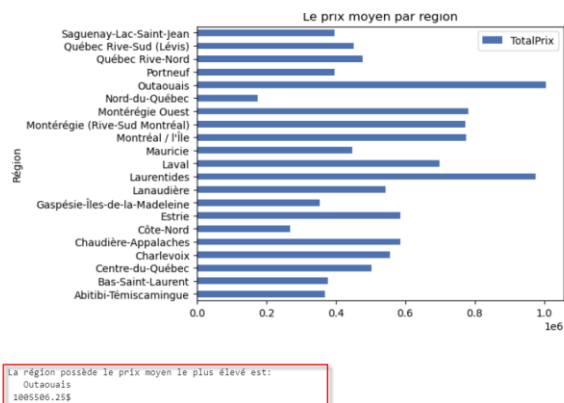
3.2 Présenter visuellement la proportion numérique de chaque région en matière de nombre d'annonces, par rapport à l'ensemble des annonces. Quelle région occupe la plus petite proportion ?



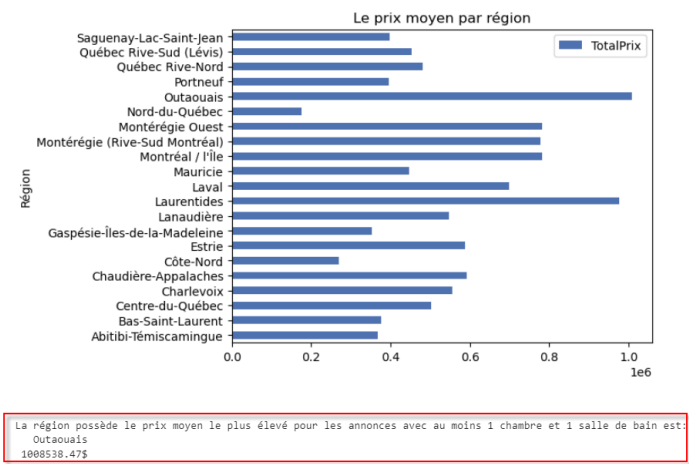
3.3 A l'aide d'un graphique différent de celui de la question précédente, comparer le nombre d'annonces de vente pour chaque région. Quelle région possède le plus d'annonces de vente ?



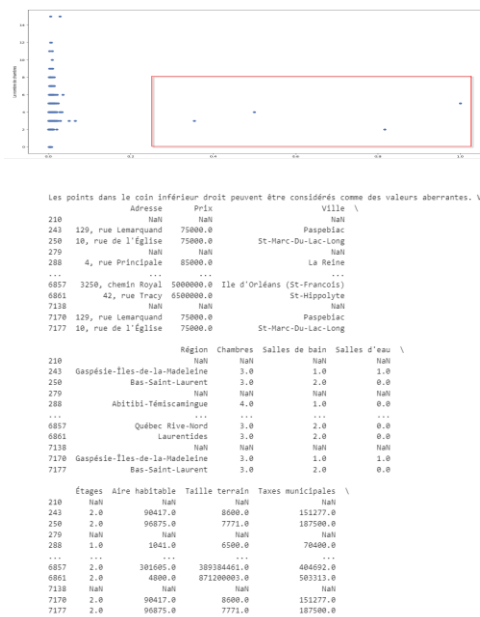
3.4 A l'aide d'un graphique, comparer le prix moyen des annonces pour chaque région. Quelle région possède le prix moyen le plus élevé ?



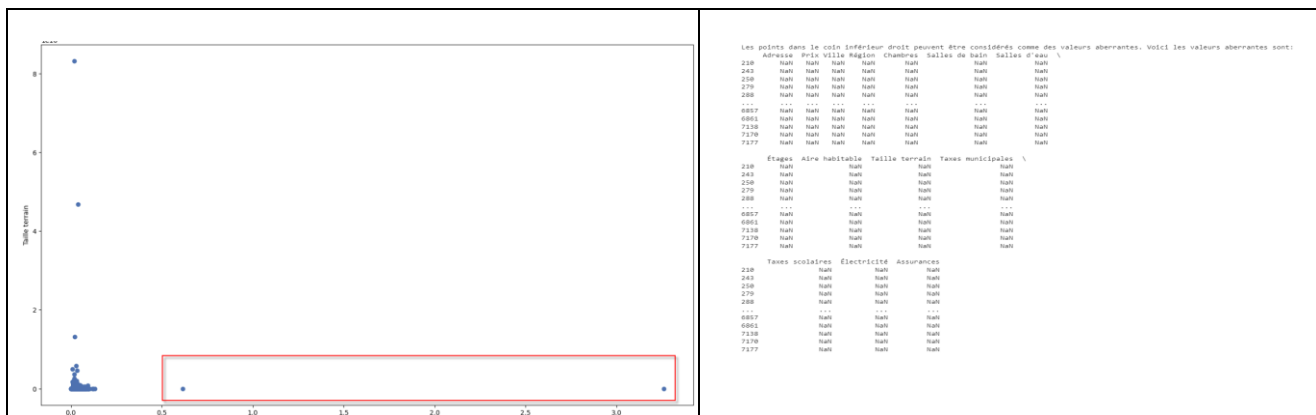
3.5 Pour ce point, on se limite aux annonces ayant au moins 1 chambre et 1 salle de bain. A l'aide d'un graphique, comparer le prix moyen de ces annonces pour chaque région. Quelle région possède le prix moyen le plus élevé pour les annonces avec au moins 1 chambre et 1 salle de bain?



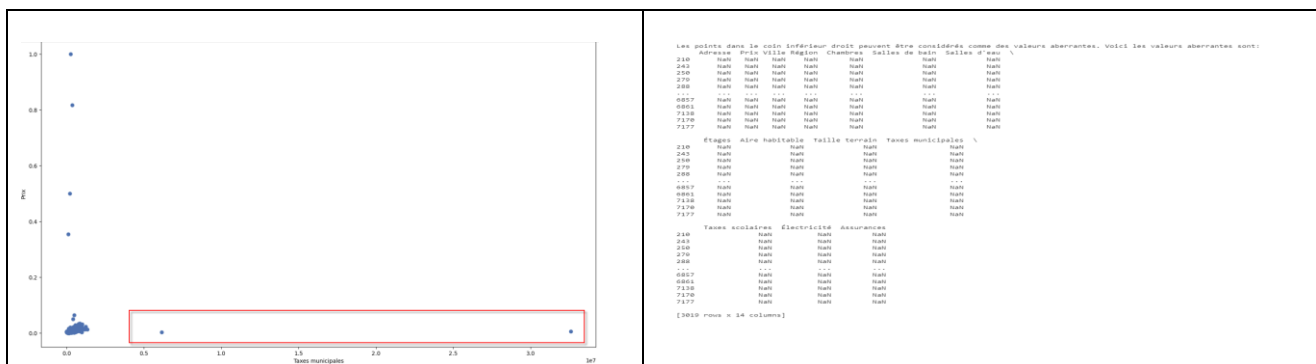
3.6 A l'aide d'un graphique, analyser la relation entre le prix des annonces et le nombre de chambres. Y a-t-il un lien quelconque ? Est-ce que la région y joue un rôle dans cette relation? Peut-on apercevoir des valeurs aberrantes ? Si oui identifiez-les : donnez toutes les valeurs des colonnes de ces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



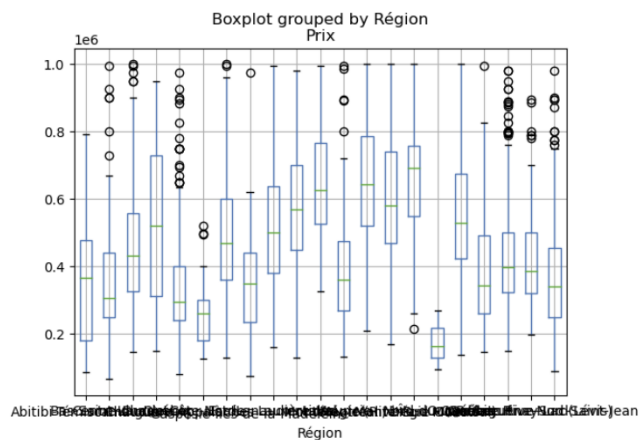
3.7 A l'aide d'un graphique, analyser la relation entre la valeur des taxes municipales annuelles des annonces et la taille du terrain. Y a-t-il un lien quelconque ? Est-ce que la région y joue un rôle dans cette relation? Peut-on apercevoir des valeurs aberrantes ? Si oui identifiez-les : donnez toutes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



3.8 A 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.

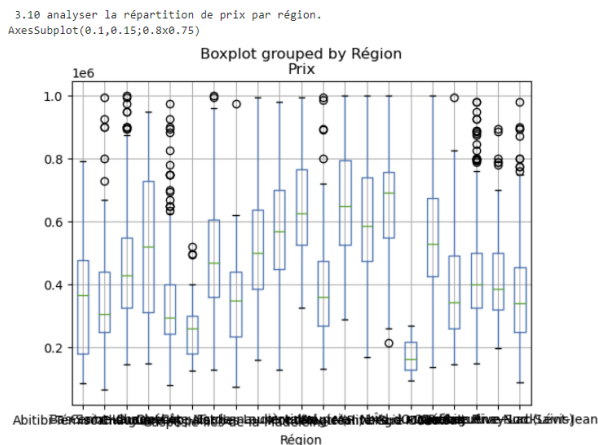


3.9 On s'intéresse pour cette question aux annonces qui ont un prix affiché de moins de 1 million de \$, pour toutes les régions. Dessiner dans un même graphique un boxplot représentant la répartition de prix par région. Analyser de manière détaillée le graphique obtenu.

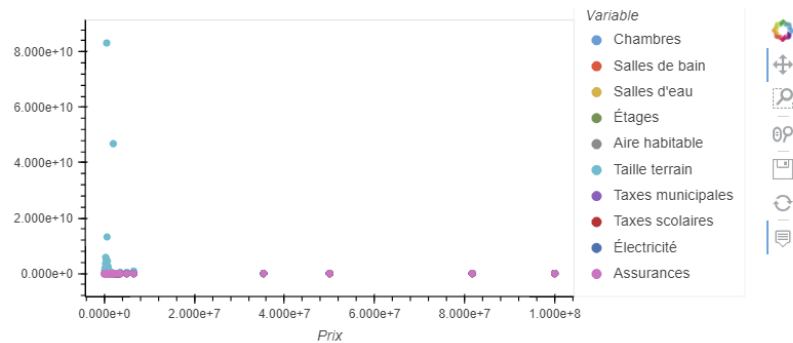


3.10 On s'intéresse pour cette question aux maisons de 2 chambres au moins et une salle de bain au moins et qui coûte moins de 1 million de \$, pour toutes les régions. Dessiner dans un même graphique un boxplot représentant la répartition de prix par régions. Analyser de manière détaillée le graphique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

Non, il n'y a pas de différences. C'est le même résultat sorti.



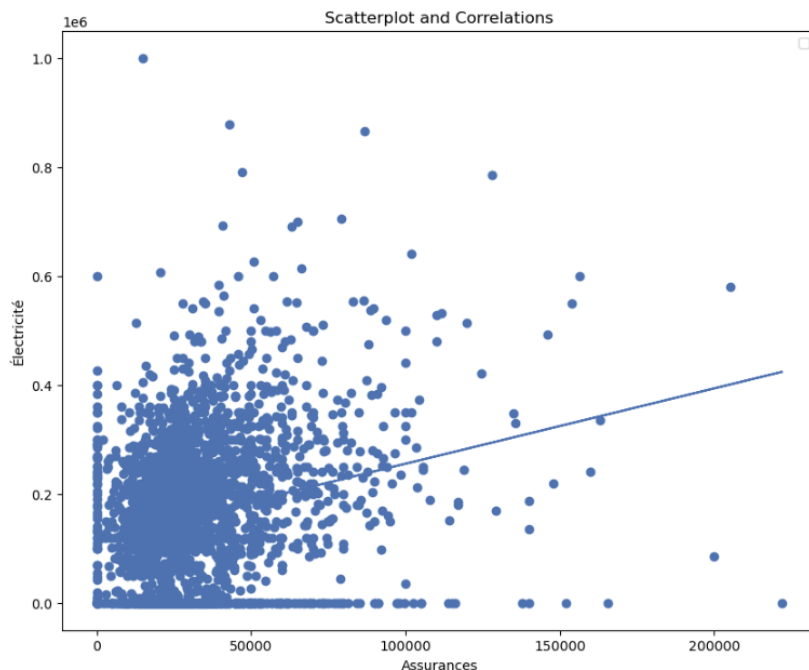
3.11 En un seul graphique, présenter une analyse bivariée de toutes les colonnes numériques de votre jeu de données. Analyser en détail le graphique obtenu.



4. PARTIE 1 : 4.4 ALGORITHMES DE RÉGRESSION (6PTS) :

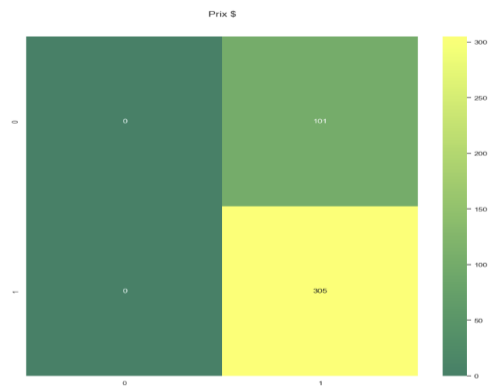
Dans cette partie, on gardera toujours 85% des données pour l'entraînement et le reste pour les tests. Remarque: Choisissez la bonne transformation pour vos données et justifiez vos choix!

4.1 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.



4.2 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.

- Transformer région en numérique.
- Ajoute une colonne de prixNum = 1 si prix > 350000, sinon prixNum=0
- x = variable région, nombre chambre,taxes scolaire
- y=prixnum
- Appliquer le modèle de la régression logistique.

[illegible]

Le prix d'une annonce est supérieur à 350000\$.

4.3 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.

```
Beta_0: 312768.593374853
Beta_1: -3684.5223907610484

=====
OLS Regression Results
=====
Dep. Variable:      Prix      R-squared:      0.006
Model:              OLS      Adj. R-squared:    0.002
Method:             Least Squares      F-statistic:    1.458
Date:              Sun, 20 Nov 2022      Prob (F-statistic): 0.158
Time:              07:09:42      Log-Likelihood:  -36622.
No. Observations:   2299      AIC:              7.326e+04
Df Residuals:       2289      BIC:              7.332e+04
Df Model:           9
Covariance Type:    nonrobust
=====
               coef      std err      t      P>|t|      [0.025      0.975]
-----
const          3.128e+05    1.7e+05    1.842    0.066    -2.03e+04    6.46e+05
Région         -3684.5224    7626.039   -0.483    0.629    -1.86e+04    1.13e+04
Chambres       -7434.7304    4.01e+04   -0.186    0.853    -8.6e+04    7.11e+04
Salles de bain  1.011e+05    7.2e+04    1.404    0.161    -4.01e+04    2.42e+05
Salles d'eau    9.899e+04    8.69e+04    1.140    0.255    -7.14e+04    2.69e+05
Étages        -1476.9061    9227.074   -0.160    0.873    -1.96e+04    1.66e+04
Aire habitable  -0.0002      0.002     -0.102    0.919    -0.004      0.004
Taille terrain  2.815e-05    4.29e-05    0.657    0.511    -5.59e-05    0.000
Taxes municipales  0.0300      0.062     0.487    0.626    -0.091      0.151
Taxes scolaires  4.3763      2.024     2.162    0.031     0.408      8.345
=====
Omnibus:          6451.927      Durbin-Watson:      2.008
Prob(Omnibus):    0.000      Jarque-Bera (JB):    170559972.695
Skew:             35.528      Prob(JB):            0.00
Kurtosis:         1335.472      Cond. No.            4.07e+09
=====

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.07e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
```

4.4 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 pi², taille du terrain: 5060 pi², 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 pi², taxes municipales: 2 324,75 \$, taxes scolaires: 65,59\$

```

Beta_0: 162075.71801195992
Beta : [ 2.11772113e+03  5.96183396e+04 -6.60002032e+04  1.55852330e+05
 -7.67731689e+02 -2.05613299e-04  3.27453034e-05  1.35684274e-02
 5.33286935e+00  6.65847669e-01  2.61315794e-01]
Mean Absolute Error: 231167.72865516314
Mean Squared Error: 93824728457.16765
Mean Root Squared Error: 306308.22459928767
Définir les caractéristiques suivantes.
Vérifiez les valeurs initialisées

```

	NumeroRegion	Chambres	Salles de bain	Salles d'eau	Étages	Aire habitable	Taille terrain	Taxes municipales	Taxes scolaires	Électricité	Assurances
0	9	3	2	1	2	1700.2	5060.00	4272.39	411.06	3584.0	110.38
1	18	2	0	0	0	0.0	8021.06	2324.75	65.59	0.0	0.00

Prédire le prix estimé de vente de chacune des deux propriétés
[386955.84438988 319812.96629179]

- Le prix estimé pour la proriété 1 est : 386955.84\$ et la propriété 2 est : 319812.96\$

4.5 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).

La ville est exclue	La ville est incluse
<pre> Beta_0: 312768.593374853 Beta_1: -3684.5223907610484 OLS Regression Results ===== Dep. Variable: Prix R-squared: 0.006 Model: OLS Adj. R-squared: 0.002 Method: Least Squares F-statistic: 1.458 Date: Sun, 20 Nov 2022 Prob (F-statistic): 0.158 Time: 07:09:42 Log-likelihood: -36622. No. Observations: 2299 AIC: 7.320e+04 Df Residuals: 2299 BIC: 7.332e+04 Df Model: 9 Covariance Type: nonrobust ===== coef std err t P> t [0.025 0.975] ----- const 3.128e+05 1.7e+05 1.842 0.066 -2.03e+04 6.46e+05 Région -3684.5224 7626.039 -0.483 0.629 -1.86e+04 1.13e+04 Chambres 7434.7304 6.01e+04 -0.126 0.853 -8.6e+04 7.11e+04 Salles de bain 1.011e+05 7.2e+04 1.404 0.161 -4.81e+04 2.42e+05 Salles d'eau 9.899e+04 8.69e+04 1.140 0.255 -7.14e+04 2.69e+05 Étages -1476.9061 3227.074 -0.458 0.649 -1.96e+04 1.66e+04 Aire habitable -0.0002 0.0002 -0.102 0.919 -0.004 0.004 Taille terrain 2.815e-05 4.29e-05 0.657 0.511 -5.50e-05 0.000 Taxes municipales 0.0300 0.062 0.487 0.626 -0.091 0.151 Taxes scolaires 4.3763 2.024 2.162 0.031 0.408 8.345 ===== Omnibus: 6451.927 Durbin-Watson: 2.008 Prob(Omnibus): 0.000 Jarque-Bera (JB): 170559972.695 Skew: 35.528 Prob(JB): 0.00 Kurtosis: 1335.472 Cond. No. 4.07e+09 ===== Notes: [1] Standard Errors assume that the covariance matrix of the errors is correctly specified. [2] The condition number is large, 4.07e+09. This might indicate that there are strong multicollinearity or other numerical problems. </pre>	<pre> Beta_0: 78411.43338095118 Beta_1: 267.8480482199357 OLS Regression Results ===== Dep. Variable: Prix R-squared: 0.008 Model: OLS Adj. R-squared: 0.003 Method: Least Squares F-statistic: 1.760 Date: Tue, 22 Nov 2022 Prob (F-statistic): 0.0628 Time: 17:11:19 Log-likelihood: -37297. No. Observations: 2299 AIC: 7.462e+04 Df Residuals: 2298 BIC: 7.468e+04 Df Model: 10 Covariance Type: nonrobust ===== coef std err t P> t [0.025 0.975] ----- const 7.841e+04 2.46e+05 0.319 0.750 -4.84e+05 5.61e+05 Région 267.8480 1.04e+04 0.026 0.979 -2.01e+04 2.08e+04 Chambres 6.015e+04 5.37e+04 1.119 0.263 -4.52e+04 1.66e+05 Salles de bain -5.339e+04 9.25e+04 -0.576 0.564 -2.35e+05 1.28e+05 Salles d'eau 1.846e+05 1.17e+05 1.582 0.114 -4.43e+04 4.13e+05 Étages -2344.6697 1.24e+04 -0.189 0.858 -2.66e+04 2.19e+04 Aire habitable -3.636e-05 0.003 -0.013 0.989 -0.005 0.005 Taille terrain 2.757e-05 5.75e-05 0.479 0.632 -8.52e-05 0.000 Taxes municipales 0.0202 0.063 0.243 0.808 -0.142 0.182 Taxes scolaires 7.8505 2.782 2.836 0.005 2.434 13.346 ville 299.6151 280.165 1.069 0.285 -249.790 849.020 ===== Omnibus: 6314.126 Durbin-Watson: 2.000 Prob(Omnibus): 0.000 Jarque-Bera (JB): 128499336.083 Skew: 33.665 Prob(JB): 0.00 Kurtosis: 1159.248 Cond. No. 4.35e+09 ===== </pre>

En ajoutant la ville, le changement est très faible, R carré est 0.008 par rapport 0.006. Une différence de 0.002. Idem pour le carré ajusté (0.003 par rapport à 0.002)

Exclure la ville

5.33286935e+00 6.65847669e-01 2.61315794e-01]

Mean Absolute Error: 231167.72865516314

Mean Squared Error: 93824728457.16765

Mean Root Squared Error: 306308.22459928767

Définir les caractéristiques suivantes.

Vérifiez les valeurs initialisées

NumeroRegion	Chambres	Salles de bain	Salles d'eau	Étages	Aire habitable	Taille terrain	Taxes municipales	Taxes scolaires	Electricité	Assurances	
0	9	3	2	1	2	1700.2	5060.00	4272.39	411.06	3584.0	110.38
1	18	2	0	0	0	0.0	8021.06	2324.75	65.59	0.0	0.00

Prédire le prix estimé de vente de chacune des deux propriétés

[386955.84438988 319812.96629179]

Inclure la ville

Beta_0: 32204.56248999706

Beta : [4.16715530e+03 8.5748397e+04 -4.27898030e+04 3.5127051e+05 -1.4906408e+03 -1.2383688e-04 3.1537816e-05 1.3980308e-02 5.4457347e+00 6.6130095e-01 2.5440800e-01 3.30817397e+02]

Mean Absolute Error: 231690.2606055952

Mean Squared Error: 9564864479.87121

Mean Root Squared Error: 305937.1723248221

Définir les caractéristiques suivantes.

Vérifiez les valeurs initialisées

NumeroRegion	Chambres	Salles de bain	Salles d'eau	Étages	Aire habitable	Taille terrain	Taxes municipales	Taxes scolaires	Electricité	Assurances	numVille	
0	9	3	2	1	2	1700.2	5060.00	4272.39	411.06	3584.0	110.38	240
1	18	2	0	0	0	0.0	8021.06	2324.75	65.59	0.0	0.00	225

Prédire le prix estimé de vente de chacune des deux propriétés

[386955.84438988 319812.96629179]

En ajoutant la ville, il y a les changements.

Le prix estimé pour la propriété 1 est : 354994.76\$ au lieu de 386955.84\$, il y a une différence de 351456.08\$
Pour la propriété 2, le prix est : 298525.25\$ au lieu de 319812.96\$, une différence de 21287.71\$

5. PARTIE 2 : 5.1 COLLECTE DE DONNÉES (6PTS) :

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.

Vous devez extraire dans un fichier .csv à remettre, l'ensemble des films qui ont les deux caractéristiques 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: https://www.imdb.com/search/title/?title_type=feature&release_date=2018-01-01,2018-12-31 . 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.

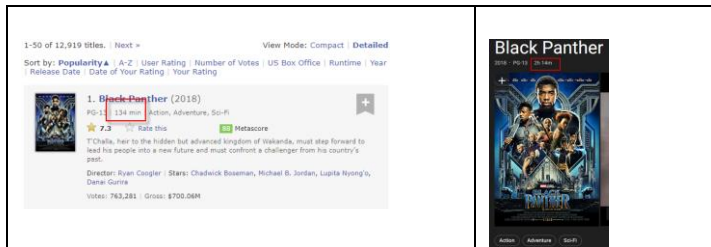
	idFilm	titreFilm	duree	genresListe	dateSortie	utilisateurNote	nbreUtilisateurVote
0	1825683	Black Panther	134.0	Action, Adventure, Sci-Fi	2018-02-16	7.3	762,371
1	7784604	Hereditary	127.0	Drama, Horror, Mystery	2018-06-08	7.3	315,114
2	4560436	Mile 22	94.0	Action, Thriller	2018-08-17	6.1	80,319
3	4154756	Avengers: Infinity War	149.0	Action, Adventure, Sci-Fi	2018-04-27	8.4	1,071,036
4	1034415	Suspiria	152.0	Drama, Fantasy, Horror	2018-11-02	6.7	79,179
...
12913	23724074	Váctimas de tratantes	89.0	Drama	NaN	NaN	NaN
12914	23731006	My Final Wife	NaN	Drama	2018-10-27	NaN	NaN
12915	23751842	Boys Club	NaN	NaN	2018-03-18	NaN	NaN
12916	23768836	El diario de una prostituta 3	NaN	Drama	NaN	NaN	NaN
12917	23769210	Perras de barrio 5	NaN	Drama	NaN	NaN	NaN

12918 rows x 7 columns

6. PARTIE 2 : 5.2 EXPLOITATION DES DONNÉES (12PTS) :

6.1 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 aberrantes, et correction du type des données (codage si c'est nécessaire). **Remarques non ordonnées:**

1. Supprimer les films dont la durée n'est pas enregistrée.
2. Convertir la durée de chaque film en minutes (entier). La durée est déjà en minute, il y a deux formats possibles. Le premier format est hh:mm, tandis que le deuxième format est en minutes, donc lors de la collecte de données, nous avons pris la durée en format minute.



3. Supprimer les films dont la durée est égale à 0.
4. Supprimer les films dont la durée est très élevée.
5. Convertir le nombre d'évaluations de chaque film nbre utilisateur note à une valeur entière.
6. Supprimer les films qui ont un nombre d'évaluations très élevée.
7. 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.
8. Convertir date sortie au format datetime.
9. Supprimer les films qui n'ont pas de date de sortie enregistrée.

Tableau final est :

idFilm	titreFilm	duration	genresListe	dateSortie	utilisateurNote	nbreUtilisateurVote	Comedy	Drama	Horror	Thriller	Action	genre_cat
297	7490386	Mar	103.0	Drama	2019-05-16	5.5	49	NaN	Drama	NaN	NaN	Drama
453	9081562	Bi bei shang geng bei shang de gu shi	105.0	Romance	2018-11-30	6.2	124	NaN	NaN	NaN	NaN	Romance
540	5233090	Yinz	89.0	Thriller	2018-10-19	5.1	120	NaN	NaN	NaN	Thriller	Thriller
579	6888362	Enthusiastic Sinners	85.0	Drama, Romance	2019-10-08	6.1	131	NaN	Drama	NaN	NaN	Romance
620	6096308	Skill	96.0	Drama, Fantasy, Mystery	2019-01-08	5.2	124	NaN	Drama	NaN	NaN	Fantasy, Mystery
...
12909	23625874	David Paquet's 2h14	NaN	Drama	2018-05-25	5.9	107	NaN	Drama	NaN	NaN	Drama
12910	23657882	Unwise King	NaN	Drama	2018-07-06	5.9	107	NaN	Drama	NaN	NaN	Drama
12911	23711076	Zoli	NaN	Drama	2018-12-07	5.9	107	NaN	Drama	NaN	NaN	Drama
12912	23722356	Efeyo	NaN	Drama	2018-12-08	5.9	107	NaN	Drama	NaN	NaN	Drama
12914	23731006	My Final Wife	NaN	Drama	2018-10-27	5.9	107	NaN	Drama	NaN	NaN	Drama

4620 rows x 13 columns

6.2 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.

	idFilm	titreFilm	duree	genresListe	dateSortie	utilisateurNote	nbreUtilisateurVote	genre_cat	dureeMinutesLog	nbreUtilisateurNoteLog
297	7490386	Mar	103.0	Drama	2019-05-16	5.5	49	Drama	6.686501	5.614710
453	9081562	Bi bei shang geng bei shang de gu shi	105.0	Romance	2018-11-30	6.2	124	NaN	6.714246	6.954196
540	5233090	Vinz	89.0	Thriller	2018-10-19	5.1	120	Thriller	6.475733	6.906891
579	6888362	Enthusiastic Sinners	85.0	Drama, Romance	2019-10-08	6.1	131	NaN	6.409391	7.033423
620	6096308	Still	96.0	Drama, Fantasy, Mystery	2019-01-08	5.2	124	NaN	6.584963	6.954196
...
12909	23625874	David Paquet's 2h14	NaN	Drama	2018-05-25	5.9	107	Drama	NaN	6.741467
12910	23657832	Unwise King	NaN	Drama	2018-07-06	5.9	107	Drama	NaN	6.741467
12911	23711076	Zoli	NaN	Drama	2018-12-07	5.9	107	Drama	NaN	6.741467
12912	23722356	Efeyo	NaN	Drama	2018-12-08	5.9	107	Drama	NaN	6.741467
12914	23731006	My Final Wife	NaN	Drama	2018-10-27	5.9	107	Drama	NaN	6.741467

4620 rows × 10 columns

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

Analyse univariée

1. Nous explorons les variables une par une.
2. La méthode d'exécution de l'analyse univariée dépendra du type de variable, qu'il soit catégoriel ou continu.

Analyse variable continue

- Mesure de la tendance centrale (moyenne, médiane, mode) de la variable.
- Mesure de la propagation (plage, IQR, variance, écart type) de la variable.
- Mesure de la forme (distribution symétrique, par exemple distribution normale, distribution asymétrique (distribution asymétrique gauche ou droite), aplatissement (forme de la distribution en termes de hauteur ou de planéité)

Analyse des variables catégorielles

- Pour les variables catégorielles, nous utiliserons la distribution de fréquence de chaque catégorie, par exemple un graphique à barres, un graphique à secteurs

Figure 1 : Nombre de répétition de la note

Text(0, 0.5, 'Nombre de repitition de la note')

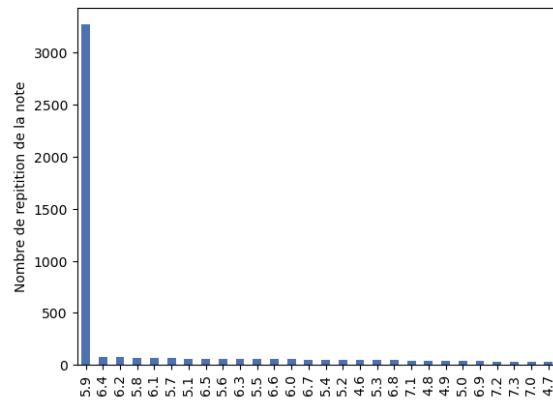


Figure 2 : Le nombre d'évaluation de chaque film de nbrUtilisateurNote

Nombre d'évaluations de chaque film de nbrUtilisateurNote

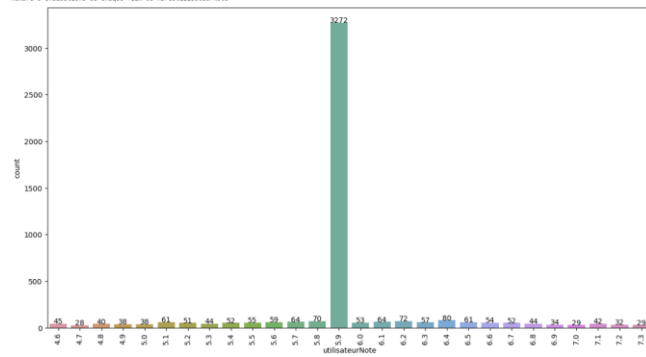


Figure 3 : La durée

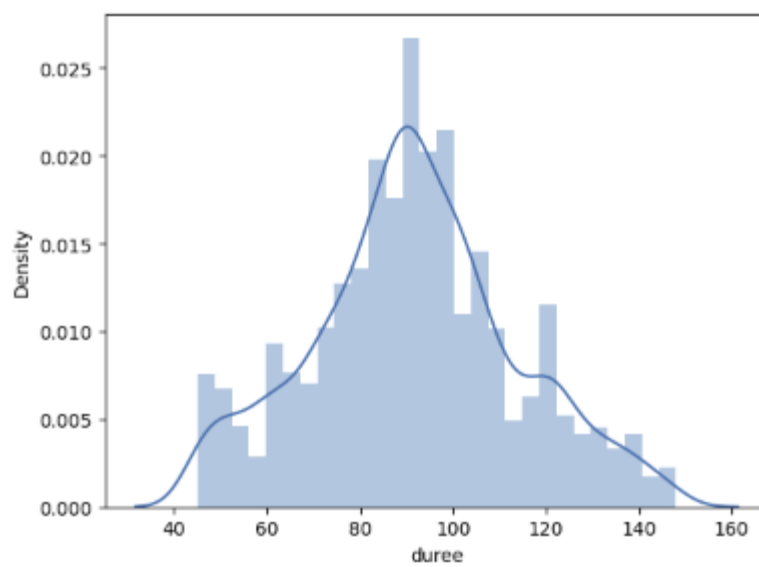


Figure 4 : Le nombre de décompte de chaque catégorie du film.

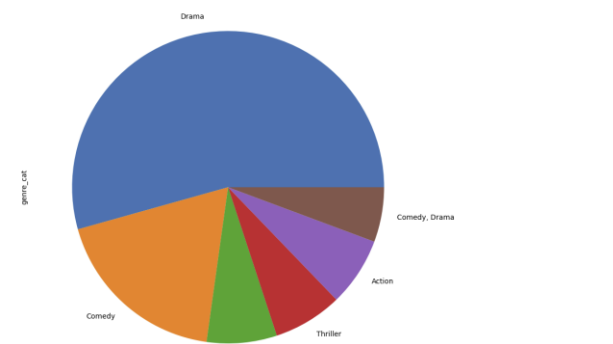


Figure 5 : Le nombre de décompte de genreListe

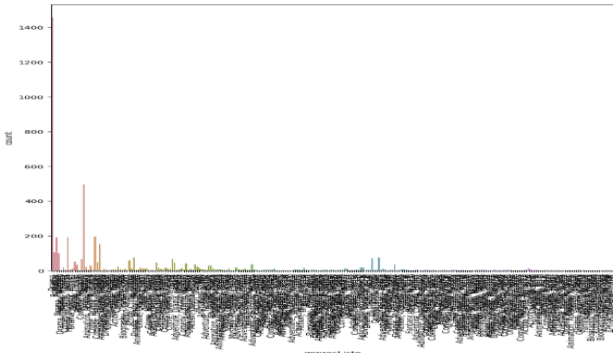


Figure 6 : La date de sortie

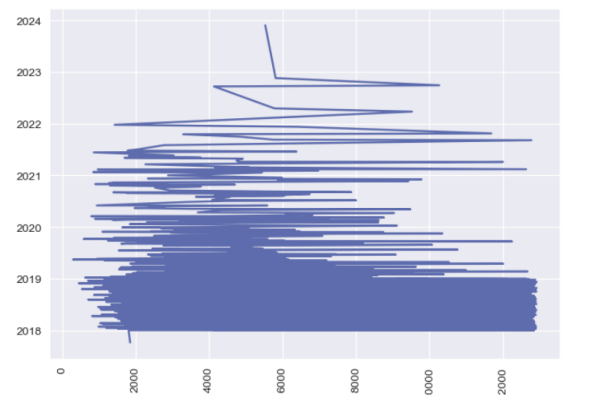


Figure 7 : la date de sortie

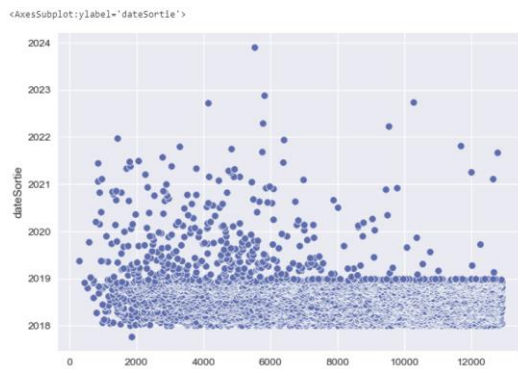
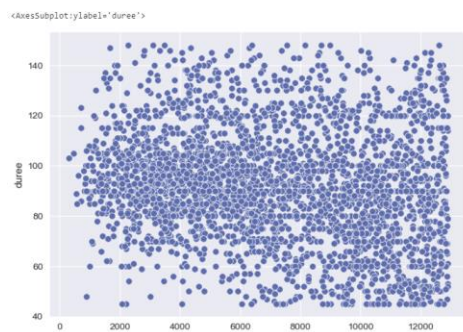


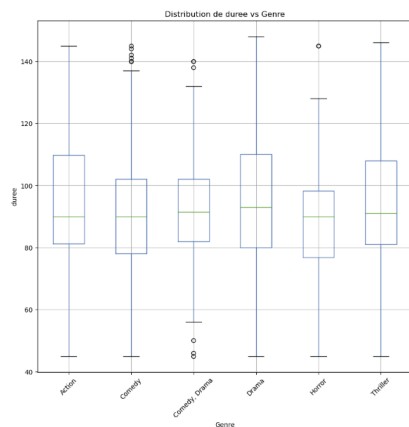
Figure 8 : La date de sortie



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

- Nous effectuons une analyse bivariée avec 2 variables pour toute combinaison de variables catégorielles et continues.
- La combinaison peut être : catégorielle et catégorielle, catégorielle et continue et continue et continue.
- Différentes méthodes sont utilisées pour aborder ces combinaisons au cours du processus d'analyse.

Figure 1 : La duree vs Genre



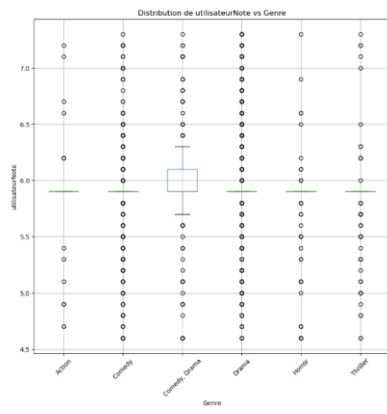


Figure 2 : Les Genres vs date de sortie

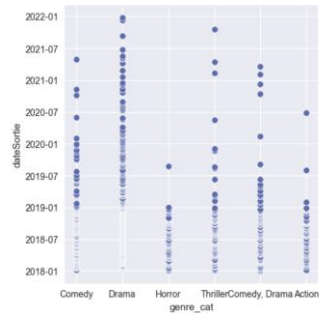


Figure 3 : Nombre de vote vs durée film

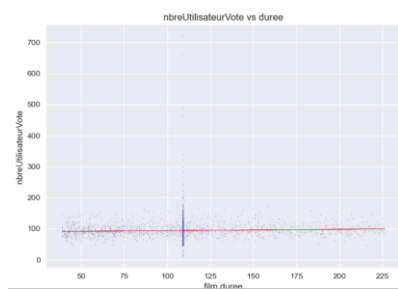


Figure 4 : Nombre de note vs genre_cat

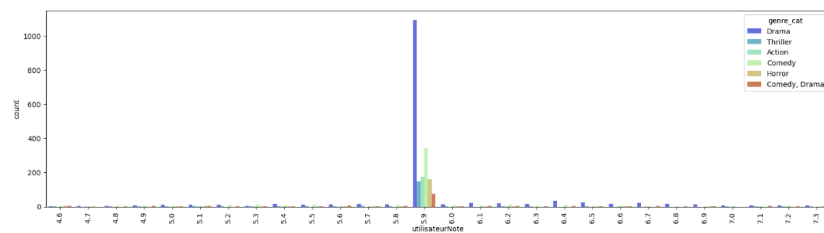


Figure 5 : UtilisateurNote vs GenreCat

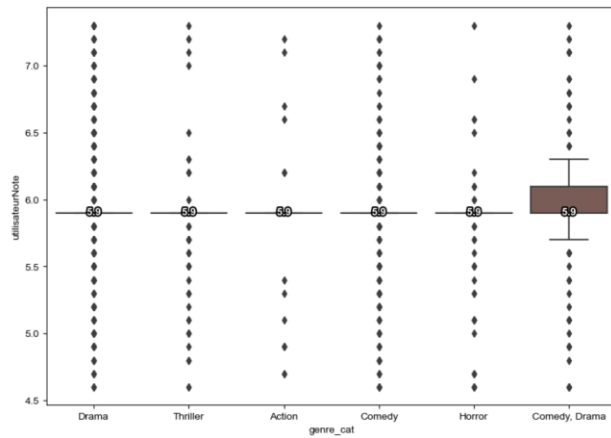


Figure 6 : durée vs GenreCat

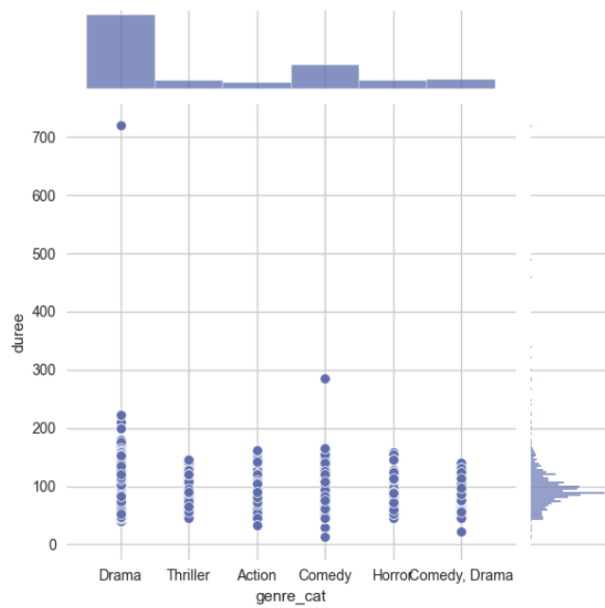
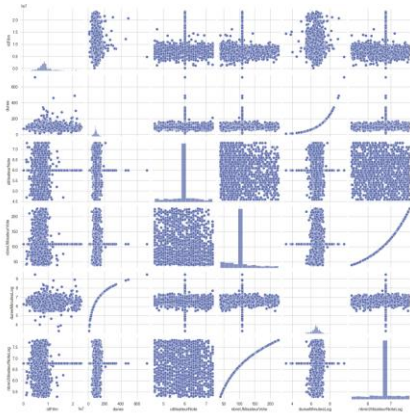


Figure 7 : pairplot



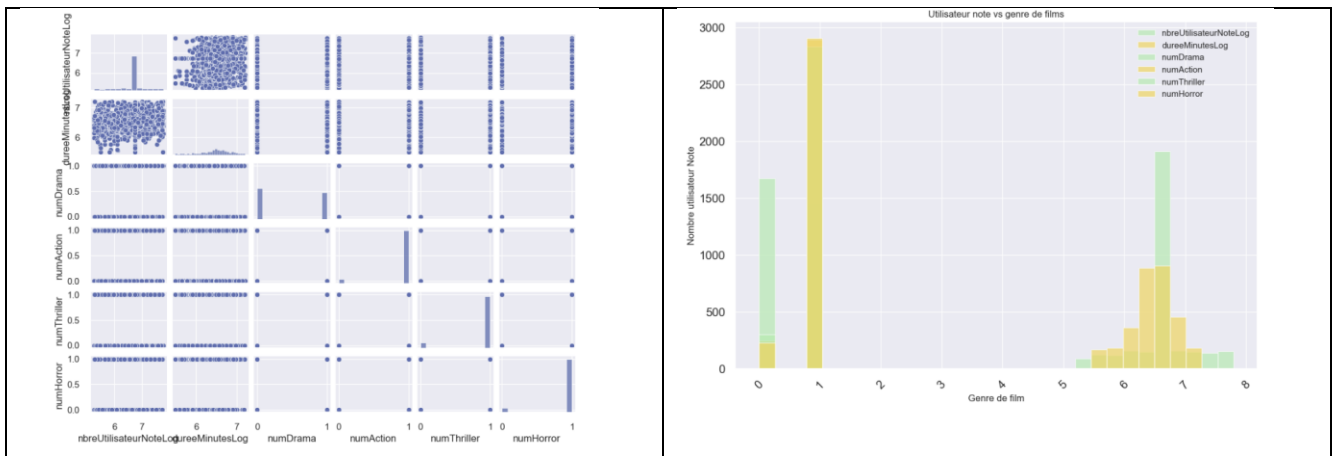
6.5 Dans cette question, on s'intéresse à prédire la note d'un film utilisateur note en fonction de 6 colonnes nbre utilisateur note log, durée minutes log, drame, action, thriller, et horreur. Concevez un modèle de régression linéaire qui permet de faire cette prédiction, vérifier les 4 conditions nécessaires pour appliquer la régression linéaire et évaluer votre modèle.

	nbreUtilisateurNoteLog	dureeMinutesLog	numDrama	numAction	numThriller	numHorror
1847	6.228819	6.870365	0	1	0	1
12088	6.741467	5.906891	1	1	1	1
11804	6.741467	0.000000	0	1	1	1
12420	6.741467	6.686501	1	1	1	1
7772	6.741467	0.000000	1	1	1	1
...
5776	6.741467	0.000000	0	1	1	1
4131	6.807355	6.894818	0	1	1	1
10271	6.741467	0.000000	1	1	1	1
5811	5.643856	6.554589	0	1	1	1
5528	6.741467	6.918863	0	1	1	1

4620 rows x 6 columns

```
Beta_0: 6.669774352921038
Beta : [-0.1297116 -0.08079081 -0.06080039 0.00996951 0.05276576 0.08834829]
Mean Absolute Error: 0.2103665677044599
Mean Squared Error: 0.15272314197557332
Mean Root Squared Error: 0.3907980833826764
```

- Les conditions requises pour que le modèle de régression linéaire multiple soit valide :
 - 1) Relations linéaires entre les variables explicatives numériques et la variable de réponse
 - 2) Les résidus doivent être presque normalement distribués.
 - 3) Variabilité constante des résidus.
 - 4) Indépendance des résidus, qui est essentiellement de l'indépendance des observations de notre échantillon.



6.6 Refaire la question précédente après avoir effectué une normalisation adéquate. Comparer les résultats.
Remarques: Pour les deux dernières questions, on regarde 80% des données pour l'entraînement et le reste pour les tests.

- On a normalisé les données en utilisant la fonction « `StandardScaler().fit_transform(x)` »

```
Beta_0: [5.91910863]
Beta : [[-0.06930368  0.01193312 -0.03747526  0.00711804  0.02028333  0.02795539]]
Mean Absolute Error: 0.26787663027498987
Mean Squared Error: 0.20042907449823263
Mean Root Squared Error: 0.4476930583538599
```

