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

Travail Pratique 2

Session : Automne 2023

Étudiant : Thomas Rivemale (RIVT02079205)

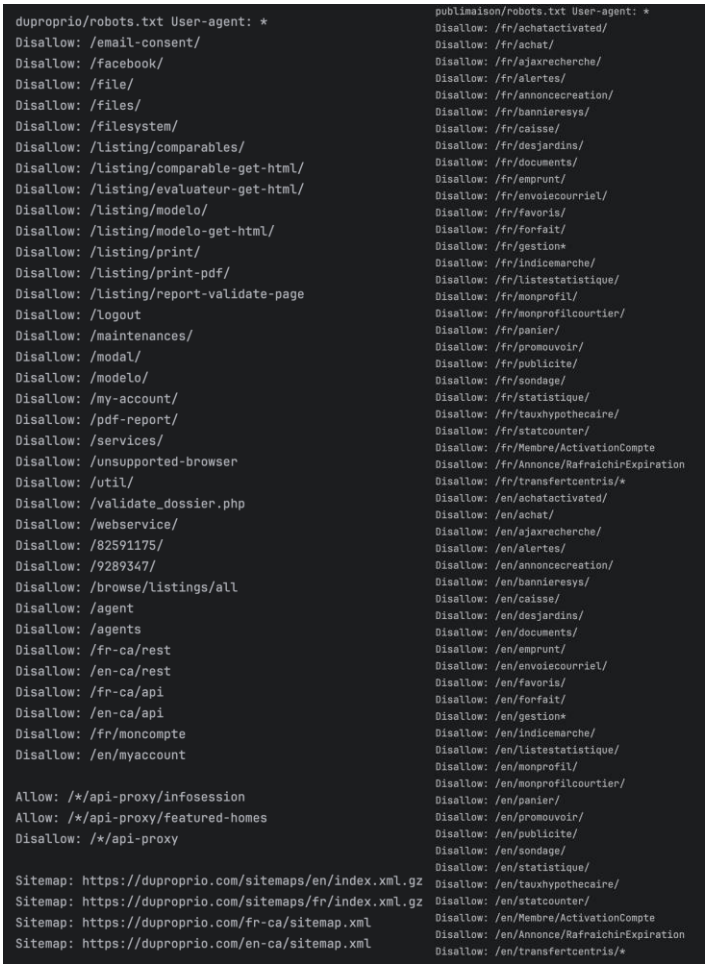
Partie 1:

1. Description du jeu de données

Le jeu de données est formé par un fichier csv.

- 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. Referez-vous a` la barre de recherche du site en question pour plus de détails.
- Région; Remarque: de même que pour les villes, referez-vous a` 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 m 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.

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



```
duproprio/robots.txt User-agent: *
Disallow: /email-consent/
Disallow: /facebook/
Disallow: /file/
Disallow: /files/
Disallow: /filesystem/
Disallow: /listing/comparables/
Disallow: /listing/comparable-get-html/
Disallow: /listing/evaluateur-get-html/
Disallow: /listing/modelo/
Disallow: /listing/modelo-get-html/
Disallow: /listing/print/
Disallow: /listing/print-pdf/
Disallow: /listing/report-validate-page
Disallow: /logout
Disallow: /maintenances/
Disallow: /modal/
Disallow: /modelo/
Disallow: /my-account/
Disallow: /pdf-report/
Disallow: /services/
Disallow: /unsupported-browser
Disallow: /util/
Disallow: /validate_dossier.php
Disallow: /webservice/
Disallow: /82591175/
Disallow: /9289347/
Disallow: /browse/listings/all
Disallow: /agent
Disallow: /agents
Disallow: /fr-ca/rest
Disallow: /en-ca/rest
Disallow: /fr-ca/api
Disallow: /en-ca/api
Disallow: /fr/moncompte
Disallow: /en/myaccount

Allow: /*/api-proxy/infosession
Allow: /*/api-proxy/featured-homes
Disallow: /*/api-proxy

Sitemap: https://duproprio.com/sitemaps/en/index.xml.gz
Sitemap: https://duproprio.com/sitemaps/fr/index.xml.gz
Sitemap: https://duproprio.com/fr-ca/sitemap.xml
Sitemap: https://duproprio.com/en-ca/sitemap.xml
```

```
publimaison/robots.txt User-agent: *
Disallow: /fr/achatactivated/
Disallow: /fr/achat/
Disallow: /fr/ajaxrecherche/
Disallow: /fr/alertes/
Disallow: /fr/annoncecreation/
Disallow: /fr/bannieresys/
Disallow: /fr/caisse/
Disallow: /fr/desjardins/
Disallow: /fr/documents/
Disallow: /fr/emprunt/
Disallow: /fr/envoiecourriel/
Disallow: /fr/favoris/
Disallow: /fr/forfait/
Disallow: /fr/gestion*
Disallow: /fr/indicemarche/
Disallow: /fr/listestatistique/
Disallow: /fr/monprofil/
Disallow: /fr/monprofilcourtier/
Disallow: /fr/panier/
Disallow: /fr/promouvoir/
Disallow: /fr/publicite/
Disallow: /fr/sondage/
Disallow: /fr/statistique/
Disallow: /fr/tauxhypothecaire/
Disallow: /fr/statcounter/
Disallow: /fr/Membre/ActivationCompte
Disallow: /fr/Annonce/RafraichirExpiration
Disallow: /fr/transfertcentris/*
Disallow: /en/achatactivated/
Disallow: /en/achat/
Disallow: /en/ajaxrecherche/
Disallow: /en/alertes/
Disallow: /en/annoncecreation/
Disallow: /en/bannieresys/
Disallow: /en/caisse/
Disallow: /en/desjardins/
Disallow: /en/documents/
Disallow: /en/emprunt/
Disallow: /en/envoiecourriel/
Disallow: /en/favoris/
Disallow: /en/forfait/
Disallow: /en/gestion*
Disallow: /en/indicemarche/
Disallow: /en/listestatistique/
Disallow: /en/monprofil/
Disallow: /en/monprofilcourtier/
Disallow: /en/panier/
Disallow: /en/promouvoir/
Disallow: /en/publicite/
Disallow: /en/sondage/
Disallow: /en/statistique/
Disallow: /en/tauxhypothecaire/
Disallow: /en/statcounter/
Disallow: /en/Membre/ActivationCompte
Disallow: /en/Annonce/RafraichirExpiration
Disallow: /en/transfertcentris/*
```

Nous décidons de ratisser le site duproprio.com. Comme nous voyons dans le robots.txt il n'est pas interdit de ratisser la liste de pagination. Seulement l'url /browse/listings/all est interdit.

2. Vous devez extraire dans un fichier .csv à remettre, l'ensemble des annonces (lancer la recherche sans aucun critère)

Pour le ratissage nous avons utilisé beautifoul soup 4. (Voir code dans notebooks)

```
2023-11-15 01:39:57,797 - INFO - Début de la collecte des données
Pages: 0%|          | 0/713 [00:00<?, ?it/s]

2023-11-15 02:36:41,433 - INFO - Fin de la collecte des données
2023-11-15 02:36:41,435 - INFO - Début de l'écriture des données
2023-11-15 02:36:41,568 - INFO - Fin de l'écriture des données
```

2. Exploration des données

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

```
Adresse               285
Prix                   68
Ville                   0
Région                  0
Chambres             1378
Salles de bain       1452
Salles d'eau         5036
Étages               2093
Aire habitable       2444
Taille terrain       1359
Taxes municipales    2170
Taxes scolaires      2564
Électricité          4126
Assurances           5241
dtype: int64
```

Voici la liste en détails des valeurs manquantes par colonnes.

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 y a plusieurs valeurs manquantes pour diverse raison :

- Données optionnelles comme les taxes et les assurances par exemple
- Données non pertinentes comme la taille du terrain ou le nombre d'étages pour un appartement
- Données non disponibles depuis la source de données

Pour les chambres, salle de bain, salle d'eau, et étages :

- Inférence : Remplacer les valeurs manquantes par le mode pourrait être une bonne solution car ce sont des variables discrètes
- Régression : Remplacer les valeurs manquantes via une régression serait une bonne solution si on arrive à démontrer une corrélation entre les variables par exemple avec la surface habitable.

Pour l'aire habitable et la taille du terrain :

- Inférence : Remplacer les valeurs manquantes par la médiane ou la moyenne pourrait être une bonne solution car ce sont des variables continue, il faudrait voir la courbe de distribution pour choisir la bonne méthode. Si la distribution est asymétrique, la médiane est plus appropriée.
- Régression : Remplacer les valeurs manquantes via une régression serait une bonne solution si on arrive à démontrer une corrélation entre les variables par exemple avec le prix.

Pour les taxes municipales, taxes scolaires, électricité et assurances :

- Inférence : Remplacer les valeurs manquantes par la moyenne ou la médiane est une solution envisageable car ce sont des variables continue, on pourrait voir s'il y a une certaine proportion avec la valeur du bien et la taille du terrain par exemple.

Pour le prix:

- Suppression: Supprimer les lignes ou le prix est manquante pourrait être une bonne solution car cela ne représente seulement que 0.94% des données.

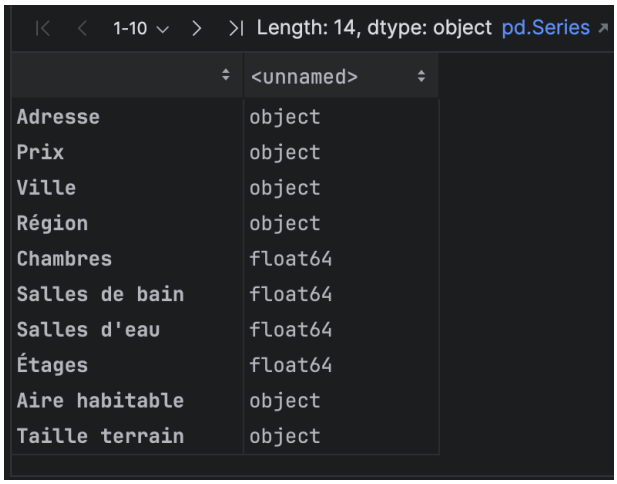
Pour l'adresse :

- Suppression: Ayant la ville et la région, on pourrait supprimer l'adresse. Puis difficilement remplaçable.

3. Combien y a-t-il de régions différentes ? et de villes différentes ?

Il y a 21 régions et 993 villes.

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



Voici les types attribuer par pandas après la lecture du fichier csv.

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

Voir notebooks.

Petite précision, j'ai décidé d'ajouter une colonne qui représente le type de bien entre terrain vide, condo et maison. Pour avoir une meilleure représentation lors de l'inférence statistique par la médiane et par des 0 dans les situations applicables. Exemple j'identifie un terrain vide, tous les caractéristique propre d'une habitation (chambre, salle de bain etc..) je remplace par 0.

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

```
Prix moyen des maisons sur l'ile de Montreal : 617155.0657939914
Prix moyen des maisons a Laval : 630601.475
Prix moyen des maisons dans les laurentides : 663159.8872545455
```

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

```
Ville la moins chere : Pointe-Aux-Trembles / Montréal-Est
Prix moyen des maisons dans la ville la moins chere : 451905.2631578947
```

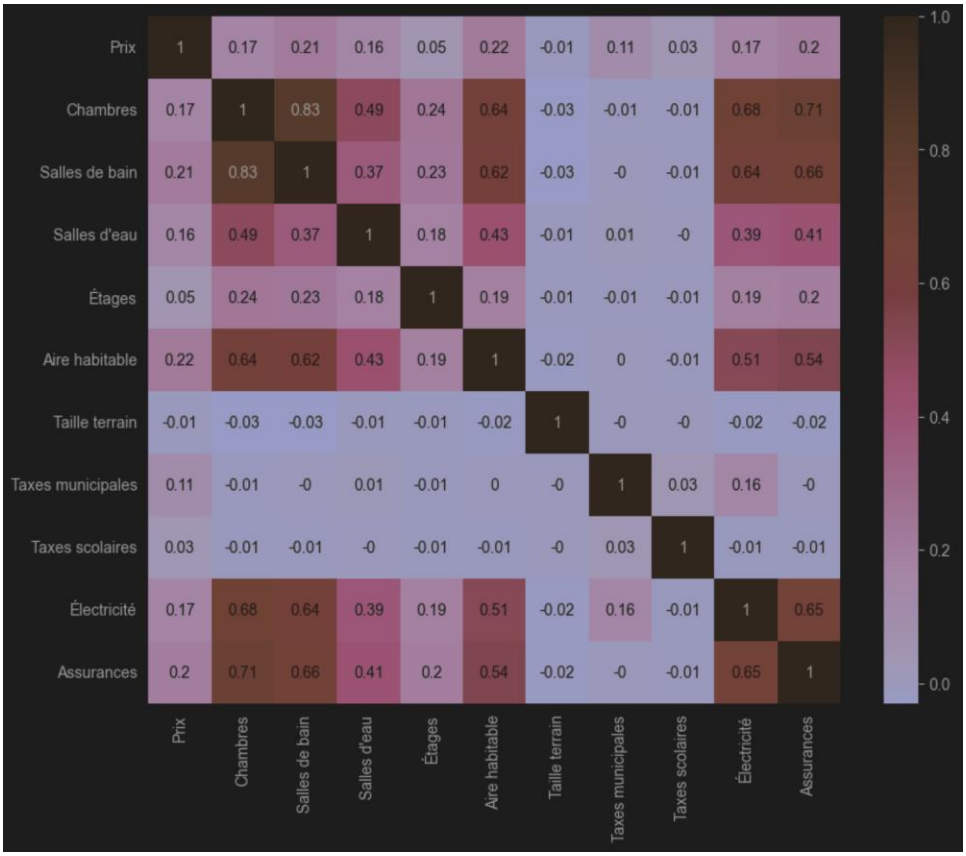
8. Pour chaque région, afficher le prix de l'item (annonce) le plus élevé et la ville où l'item se situe. Ici on ne fait pas de différence si c'est un condo/appartement, maison, terrain vide, etc. A quel région/ville revient la palme d'or de l'item le plus cher ? Donner toutes les caractéristiques (valeurs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

```
Item le plus cher par région :
                          Région                            Ville                        Adresse       Prix
200            Abitibi-Témiscamingue                      Senneterre                 550, 10e Avenue  1200000.0
3958               Bas-Saint-Laurent  Rimouski (Ste-Odile-Sur-Rimouski)          137, chemin des Pointes  1750000.0
222                 Centre-du-Québec                   St-Pie-De-Guire                   195, 6e Rang  5200000.0
3033                       Charlevoix                 Les Éboulements    438, chemin Catherine-Delzenne  2585000.0
4318            Chaudière-Appalaches                       St-Julien             547, chemin Lehoux  6500000.0
895                       Côte-Nord                       Sept-Iles             600, avenue Cartier  1200000.0
4304                         Estrie                          Granby  971-993, rue Henry-Carleton-Monk  3300000.0
6544   Gaspésie-Îles-de-la-Madeleine                           Percé                 150, route 132  2300000.0
6125                     Lanaudière         Terrebonne (Terrebonne)           1355, boulevard Moody  2775000.0
6007                     Laurentides                     St-Hippolyte                   42, rue Tracy  6500000.0
6760                           Laval                     Duvernay-Est  5967, Rang du Bas-Saint-François  4000000.0
7048                        Mauricie      Shawinigan (Shawinigan-Sud)               3980, 105e Avenue  3500000.0
6738               Montréal / l'Île  Pointe-Aux-Trembles / Montréal-Est  11788-11820, rue Notre-Dame Est  5750000.0
207   Montérégie (Rive-Sud Montréal)                         Beloeil                147, Rue Brillon  3300000.0
4619               Montérégie Ouest            Vaudreuil-Sur-Le-Lac          127, rue des Aubépines  1999900.0
4899                 Nord-du-Québec                          Chapais      128, chemin du Lac Opémiska   425000.0
1480                      Outaouais                 Gatineau (Aylmer)           125, chemin Rivermead  6500000.0
6594                       Portneuf           St-Marc-Des-Carrières          1818, avenue Principale  1850000.0
6494               Québec Rive-Nord      Ile d'Orléans (St-Francois)              3250, chemin Royal  5000000.0
6641        Québec Rive-Sud (Lévis)                      St-Romuald     731-751 rue de Saint-Romuald  5365000.0
7041        Saguenay-Lac-Saint-Jean             St-Henri-De-Taillon         chemin du Domaine-Renaud  2700000.0
```

```
Item le plus cher :
Adresse                125, chemin Rivermead
Prix                                6500000.0
Ville                     Gatineau (Aylmer)
Région                              Outaouais
Chambres                                    5
Salles de bain                              2
Salles d'eau                                1
Étages                                      3
Aire habitable                         541.81
Taille terrain                      19424.928
Taxes municipales                     2870.31
Taxes scolaires                        281.0
Électricité                           1956.0
Assurances                            1142.36
Type                                   Maison
Name: 1480, dtype: object
```

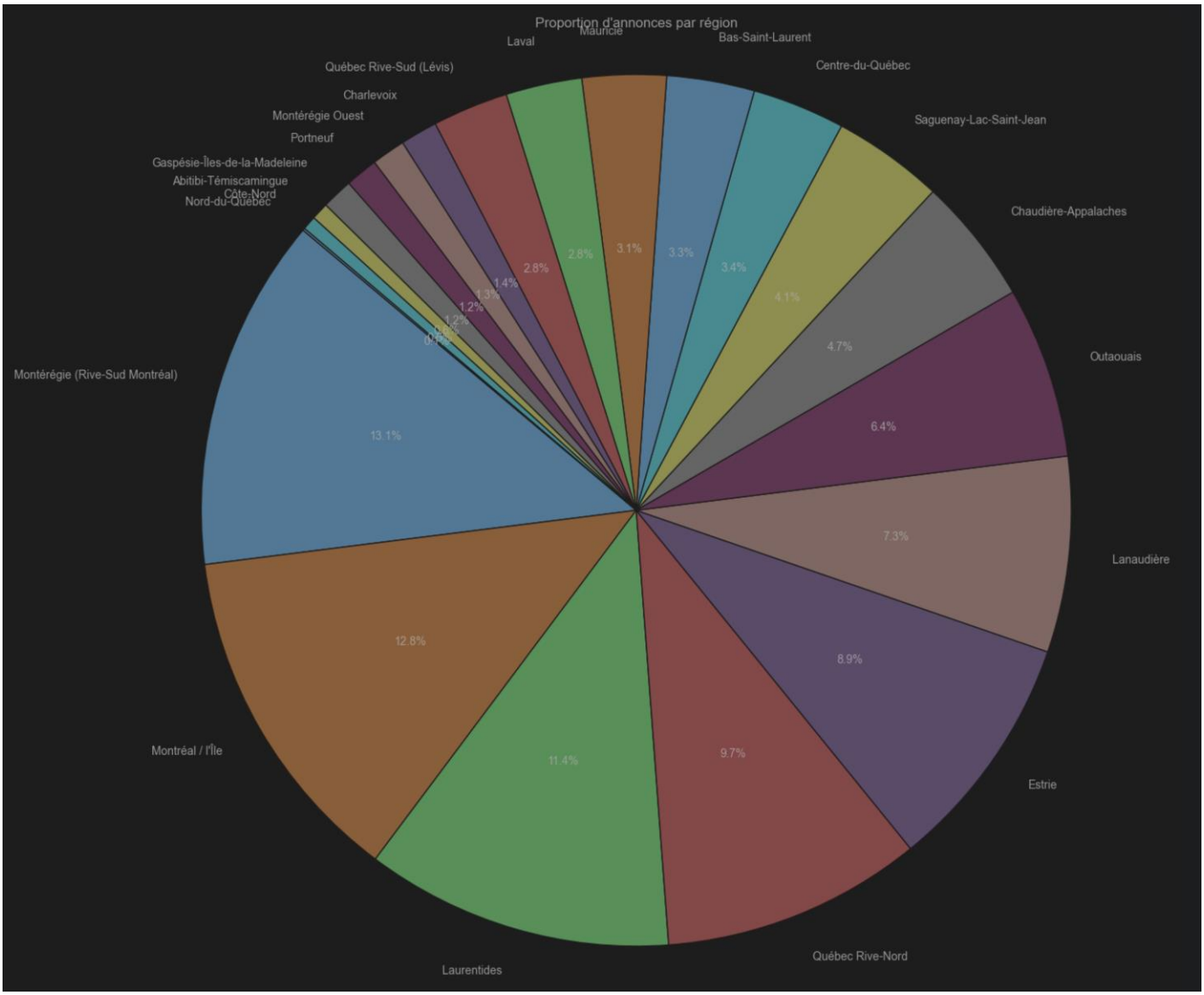
3. Visualisation et analyse des données

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 ?



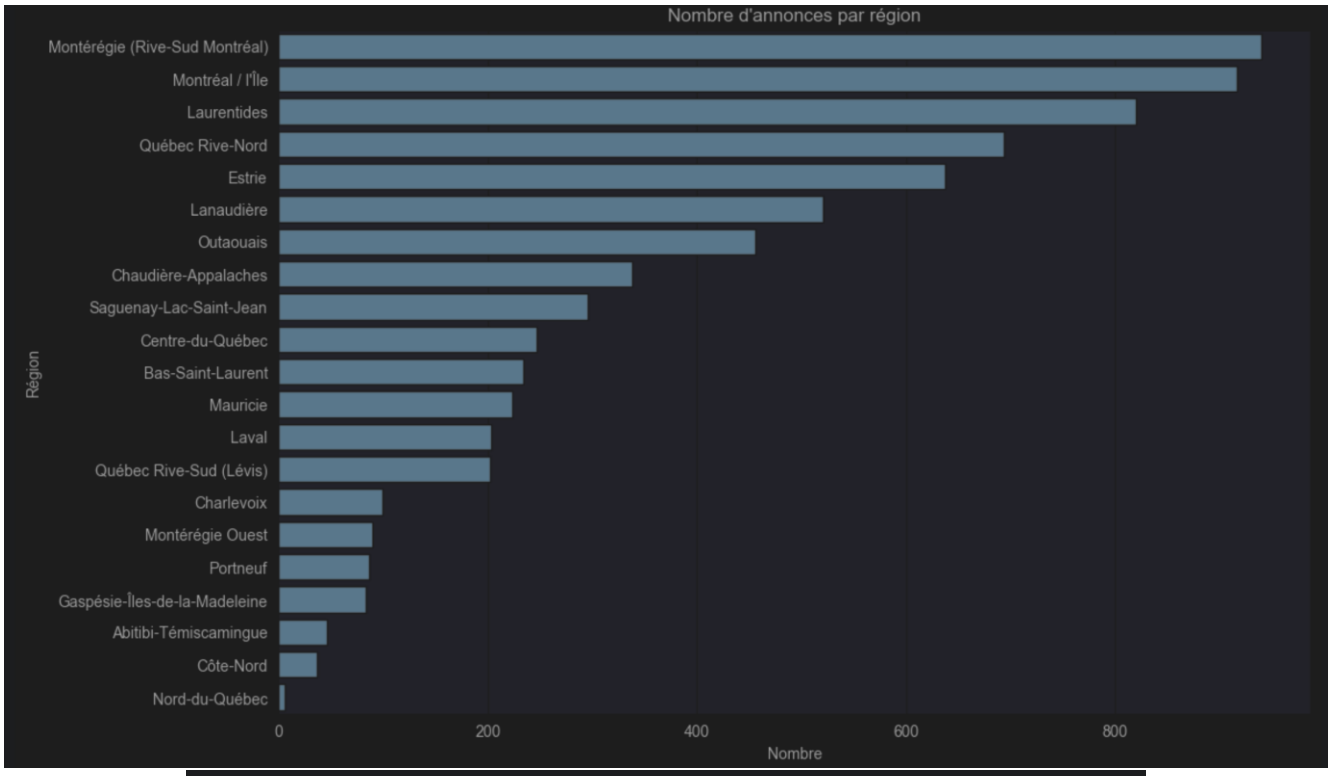

```
Correlation superieur a 0.7 :
 Chambres        Salles de bain    0.83
                 Assurances        0.71
Salles de bain   Chambres          0.83
Assurances       Chambres          0.71
dtype: float64
```

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 ?



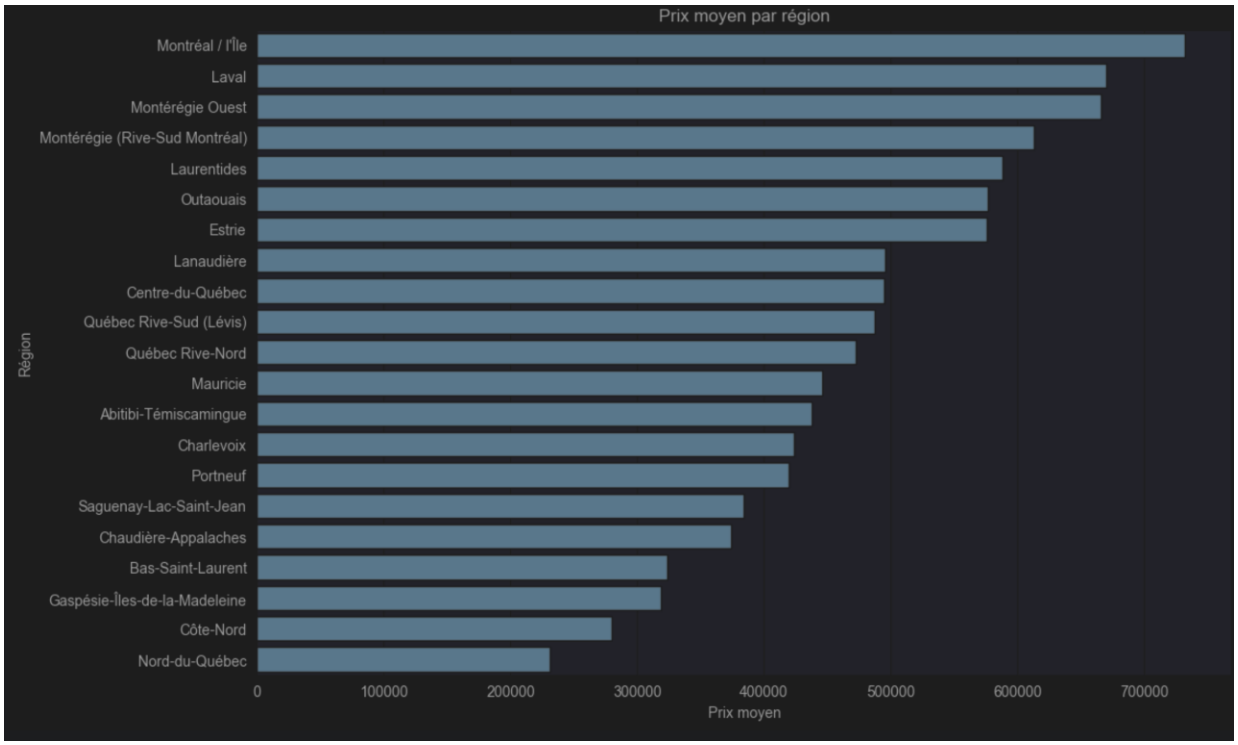

```
Région avec la plus petite proportion : Nord-du-Québec
```

3. A l'aide d'un graphique diffèrent 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 ?



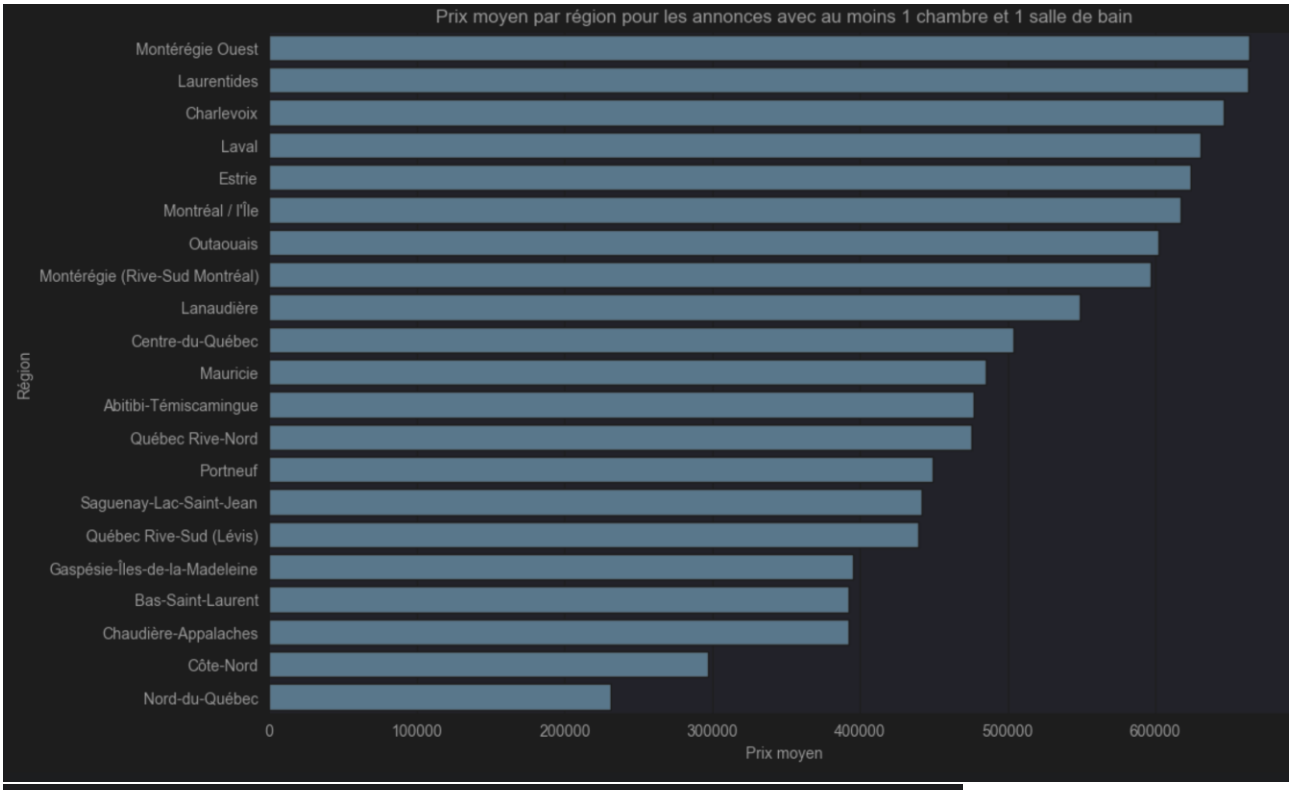

```
Région avec le plus d'annonces : Montérégie (Rive-Sud Montréal)
```

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 ?



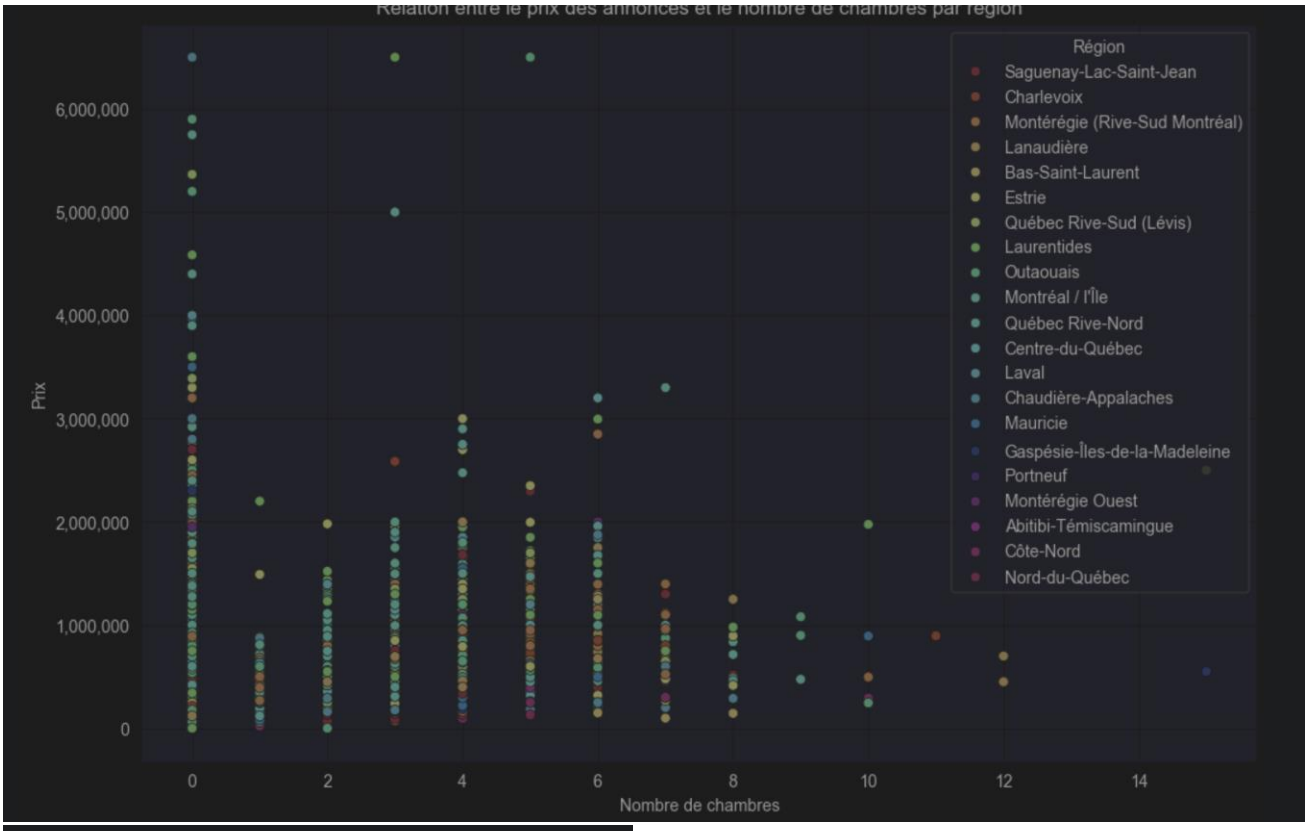

```
Région avec le prix moyen le plus élevé : Montréal / l'Île
```

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?




```
Région avec le prix moyen le plus élevé pour les annonces avec au moins 1 chambre et 1 salle de bain :
 Montérégie Ouest
```

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.




```
Valeurs aberrantes prix:
[2000000.   1360000.   1250000.   1294000.   2200000.   3300000.
 5200000.   1245000.   1485000.   2699000.   3000000.   1999999.
 1595000.   1325000.   1400000.   1395000.   1590000.   1600000.
 1375000.   1688000.   1299000.   1549000.   1259900.   1995000.
 1658000.   1279000.   1365000.   1249000.   1495000.   2500000.
 1449000.   1399000.   1550000.   1489000.   6500000.   1499000.
 1625000.   2999000.   2550000.   2850000.   1719000.   1750000.
 2295000.   1399999.   2212800.   1669000.   1390000.   1441270.
 1450000.   1229000.   1884014.84 1575000.   1390737.   1354206.6
 1396719.75 1349000.   1598000.   1354361.   1299900.   1273000.
 1434000.   4400000.   1850000.   1385000.   1329999.   1599000.
 1300000.   1220000.   1648000.   2585000.   1295000.   1695000.
 5900000.   1520000.   2299900.   1845000.   1275000.   1548000.
 1725000.   1220400.   1650000.   1279900.   1700000.   3977170.
 1990000.   1398900.   1624900.   1298000.   2199000.   1975000.
 1234567.89 1350000.   1780000.   1270000.   1777000.   1288000.
 1945000.   1518000.   1960000.   1490000.   2125000.   1448000.
 1999900.   1235000.   2995000.   1309000.   1665000.   1225000.
 1599999.   1679000.   1320000.   1888000.   3600000.   1230000.
 1959000.   1348000.   1293000.   1858000.   1589000.   1475000.
 1649000.   2450000.   1950000.   2900000.   2600000.   2750000.
 1699000.   1500000.   1680000.   1850999.   2069000.   1398000.
 3389000.   1335000.   1349500.   3900000.   2775000.   3200000.
 1875000.   2100000.   5199900.   1290000.   1675000.   2362690.
 1330000.   1988000.   1380000.   5000000.   1915000.   2300000.
 2800000.   5365000.   1585000.   1800000.   5750000.   4000000.
 2400000.   1259000.   2475000.   1299999.   2920000.   1588000.
 1498000.   1980000.   2700000.   3500000.   1899000.   2350000.
 4585000.   2999999.   1799999.   1998000.   1790000.   1949000.
 1470000.   1900000.   1499999.   1999000.  ]
Valeurs aberrantes chambres:
<IntegerArray>
[8, 15, 9, 10, 11, 12]
Length: 6, dtype: Int64
```

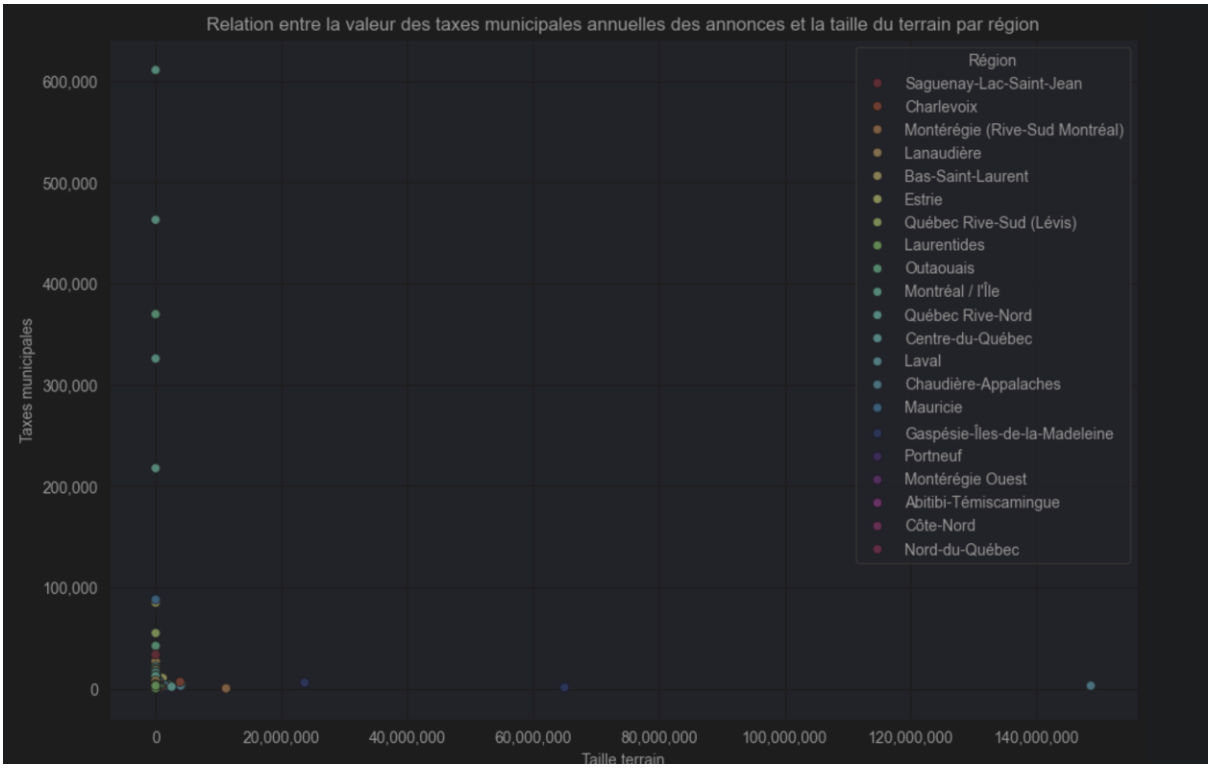
```
Prix
Q1 : 320000.0
Q3 : 679000.0
IQR : 359000.0
Max : 1217500.0
Min : -218500.0

Chambres
Q1 : 0.0
Q3 : 3.0
IQR : 3.0
Max : 7.5
Min : -4.5
```

D’après ce graphique on ne peut pas conclure que le nombre de chambres par lui-même n’influe pas directement sur le prix de l’annonce, car on n’identifie pas de tendances particulières.

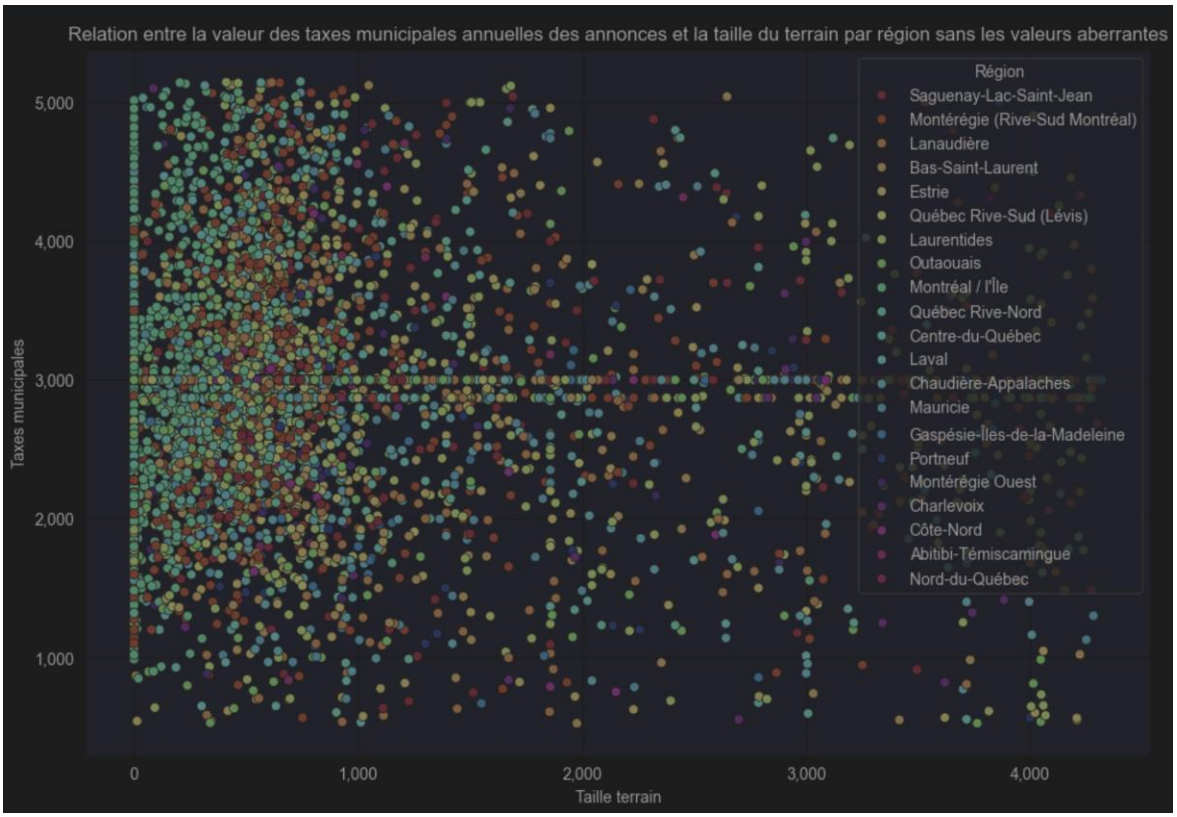
Par contre aux vues des couleurs des régions, on remarque que certaines d’entre elle ont une valeur plus élevée par rapport à certaines, cela pourrait signifier que la région a une influence sur le prix d’un bien immobilier.

Concernant les valeurs aberrantes, on identifie dans les deux cas. Dans les prix ont observé des valeurs aberrantes sur le graphique, les chambres élevées ainsi que des prix élevés. Pour l'identifier plus précisément, on a utilisé la méthode écart interquartile. Pour les chambres l’IQR nous donne 7.5 chambres donc nous pouvons dire dans le contexte immobilier 8 chambres. Et pour le prix on a 1 217 500 $ ce qui peut paraitre aberrant sur l’ensemble des régions, mais pour cette région cela peut ne pas être aberrants.

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.




```
Taxes municipales
Q1 : 2258.19
Q3 : 3415.16
IQR : 1156.9699999999998
Max : 5150.615
Min : 522.7350000000004

Taille terrain
Q1 : 248.16
Q3 : 2043.9
IQR : 1795.74
Max : 4737.51
Min : -2445.4500000000003
```

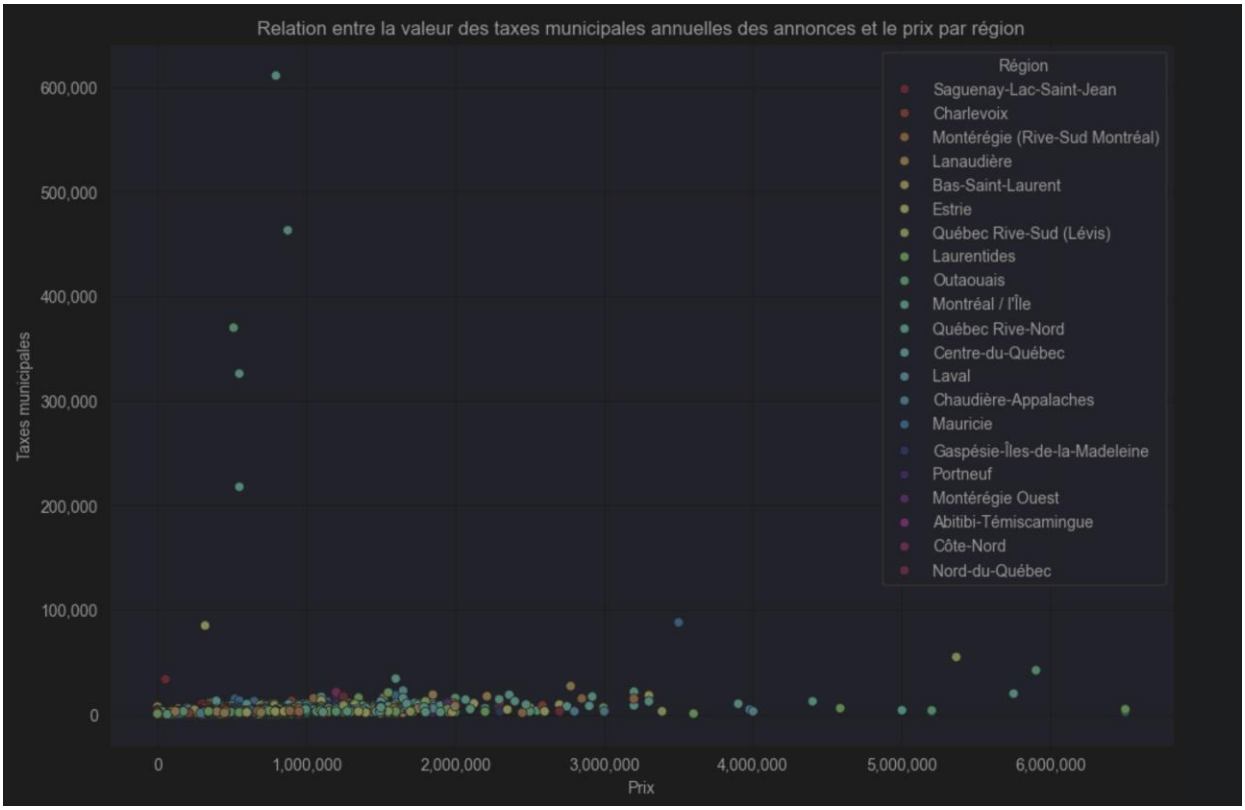


Le premier graphe nous montre des valeurs aberrantes qui étale beaucoup l'échelle pour voir des relations. Juste un amas de points à l'origine. On a donc fait un deuxième graphique sans valeurs aberrantes en utilisant l'écart interquartile pour essaye de voir des relations.

Comme pour la question précédente on ne voit pas de relation linéaire entre ces deux variables. Visuellement c'est plus la région qui fait office d'indicateur en évolution du prix que la taille du terrain.

Pour le graphique des valeurs aberrants on a enlevé les tailles de terrain supperieur a 4737.51 m2 et les taxes municipales supérieure à 5150.62$. Ces valeurs aberrantes peuvent être des erreurs des saisis ou, compte tenu de notre contexte immobilier, des biens exceptionnels.

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.

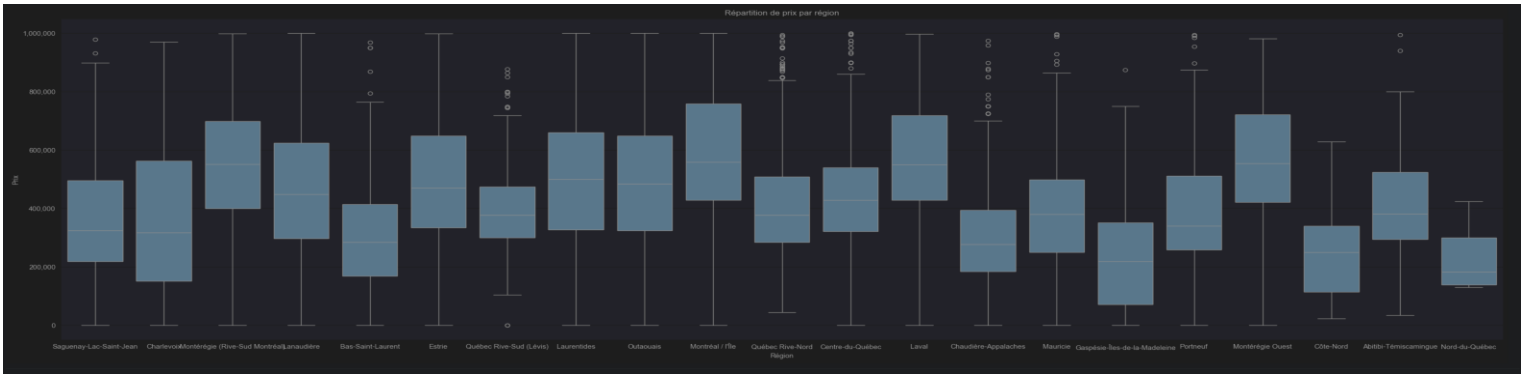



```
Taxes municipales
Q1 : 2258.19
Q3 : 3415.16
IQR : 1156.9699999999998
Max : 5150.615
Min : 522.7350000000004

Prix
Q1 : 320000.0
Q3 : 679000.0
IQR : 359000.0
Max : 1217500.0
Min : -218500.0
```

Cette fois-ci encore nous avons dû identifier les valeurs aberrantes et les enlever pour voir s'il y avait une tendance. Car les valeurs aberrantes rendent une echelle trop grande pour voir.

Dans le graphique sans valeurs aberrantes, on remarque une certaine relation linéaire, quand le prix augmente la taxe municipale augmente. Cela ne va pas forcément dire qu'il y a une forte corrélation ça peut être le résultat d'autre corrélation. Par exemple le prix augmente en fonction de la région ainsi de ce fait les taxes augmentes car elles sont plus chères dans cette région.

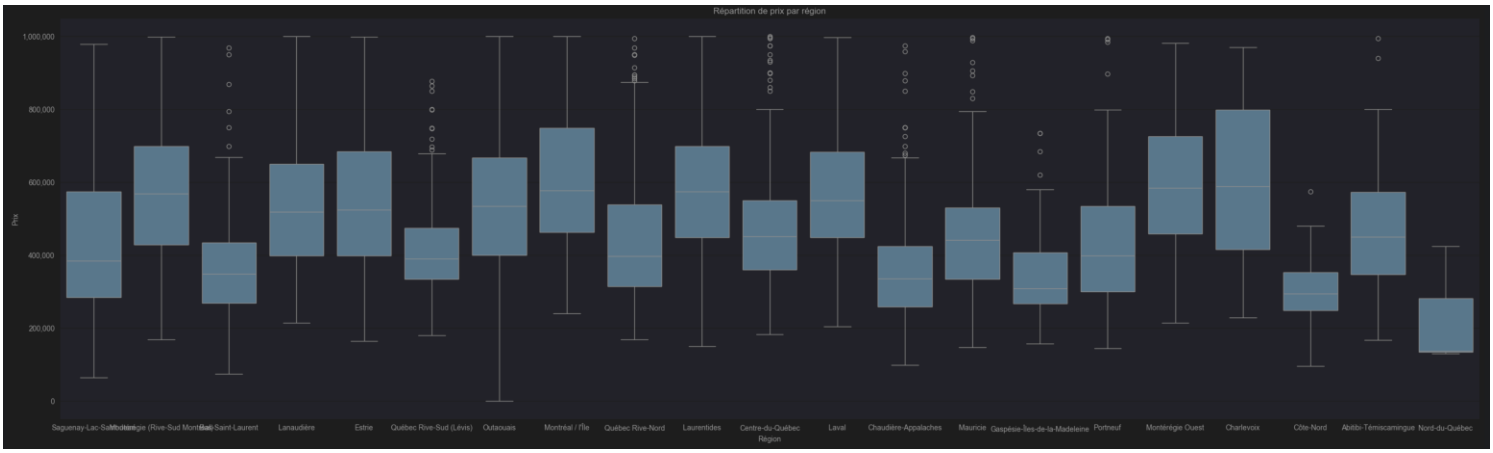
9. On s'intéresse pour cette question aux annonces qui ont un prix affiche de moins de 1 million de $, pour toutes les régions. Dessiner dans un même graphique un boxplot représentant la répartition de prix par région. Analyser de manière détaillée le graphique obtenu.



A partir de ce graphique on peut établir des tendances pour certaines régions et voir qu'elles sont les régions avec des prix élever ou moins chères. On peut aussi voir que dans certaines régions il y

a des bien surévaluer (les valeurs aberrantes) si l'entrée n'est pas une erreur de saisie. Ce graphique permet de faire des comparaisons entre les régions. On peut aussi évaluer les distributions des régions, par exemple les Laurentides visuellement on pourrait voir une distribution symétrique alors que Bas Saint Laurent une asymétrie à gauche et pour l'ile de Montréal une asymétrie à droite. Pour le cas de Montréal cela voudrai dire que la distribution de maison se rapproche plus de 1Millions que le Bas-Saint-Laurent. On remarque aussi que la médiane de prix change en fonction de la région.

10. On s'intéresse pour cette question aux maisons de 2 chambres au moins et une salle de bain au moins et qui coute 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.



Dans les différences on remarque que les médianes, la distribution, le 1er et 3eme quartile et l'apparition de valeurs aberrantes. Par exemple pour Montréal le 1er quartile a augmenté, pour la région Québec Rive-sud une valeur aberrante a disparu, pour la région Saguenay la médiane s'est rapprocher des 400 000$ et la distribution de la région de Charlevoix est passe d'asymétrique à gauche a, visuellement, symétrique.

Le fait de changer les caractéristiques d'observations peut faire apparaitre de nouvelles tendances/distribution.

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.

Pour réaliser l'analyse bivarié on a fait un pairplot , voir dans le zip car trop gros pour mettre dans le rapport.

Dans notre cas avec 11 colonnes numériques on a un pairplot composer des 121 graphiques. Donc j'ai fourni quelque observation pour exemple.

Grace au pairplot qui offre tous les scatterplot possible pour chaque combinaison de variables, on peut avoir une visualisation sure:

- Relation linéaire

On peut examiner chaque scatterplot pour voir s'il y a des relations linéaires. Par exemple on peut voir qu'il y a une certaine relation linéaire dans la combinaison Chambre-Salle de bain.

- Valeurs Aberrantes

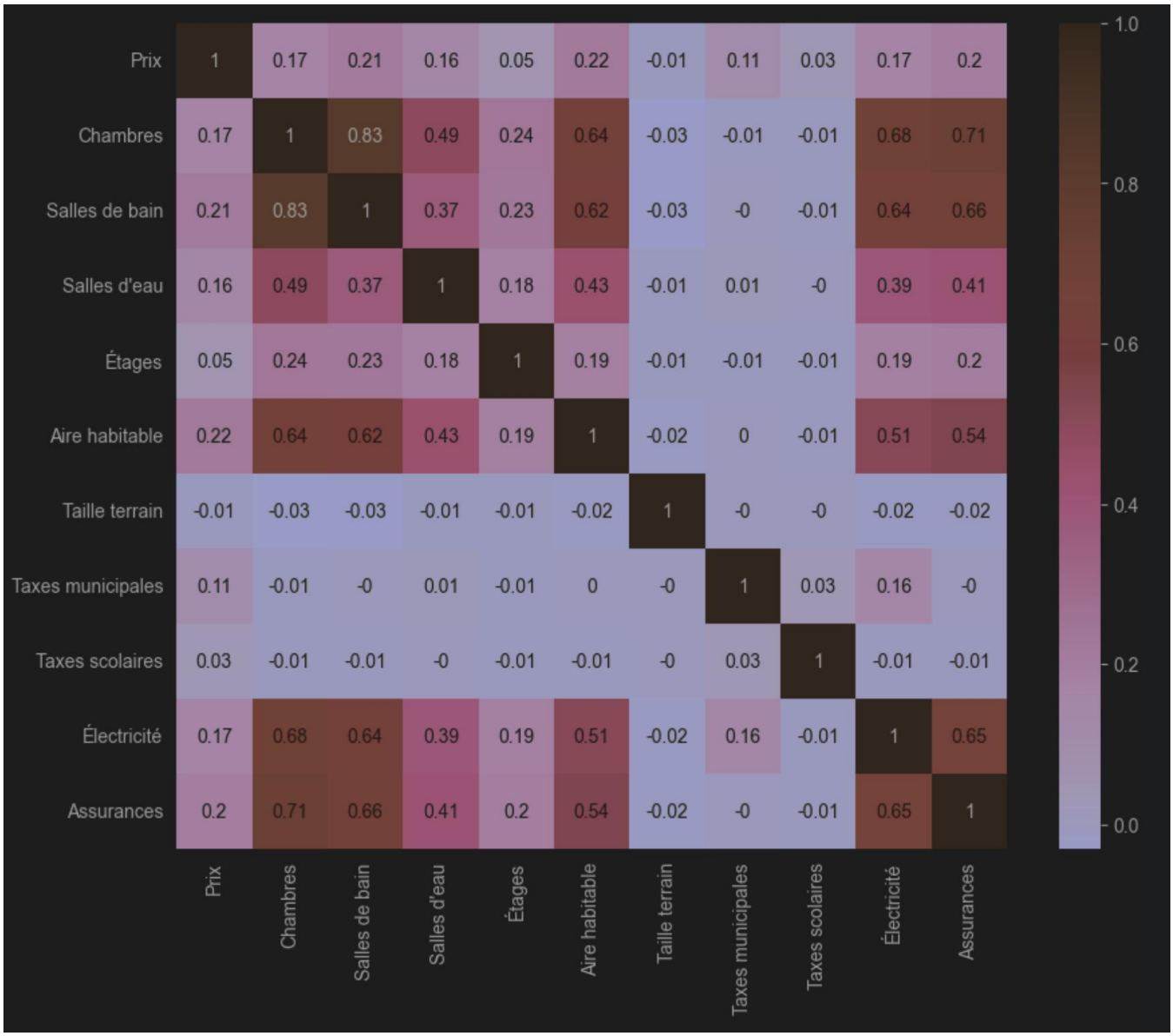
On observe des valeurs aberrantes dans certaine combinaison comme dans Prix-Chambre, Prix-Salle de bains.

- Corrélation

On peut aussi visuellement établir des corrélations visibles entre certaines variables, comme Prix-Assurance on peut voir qu'il y a une corrélation positive entre elles.

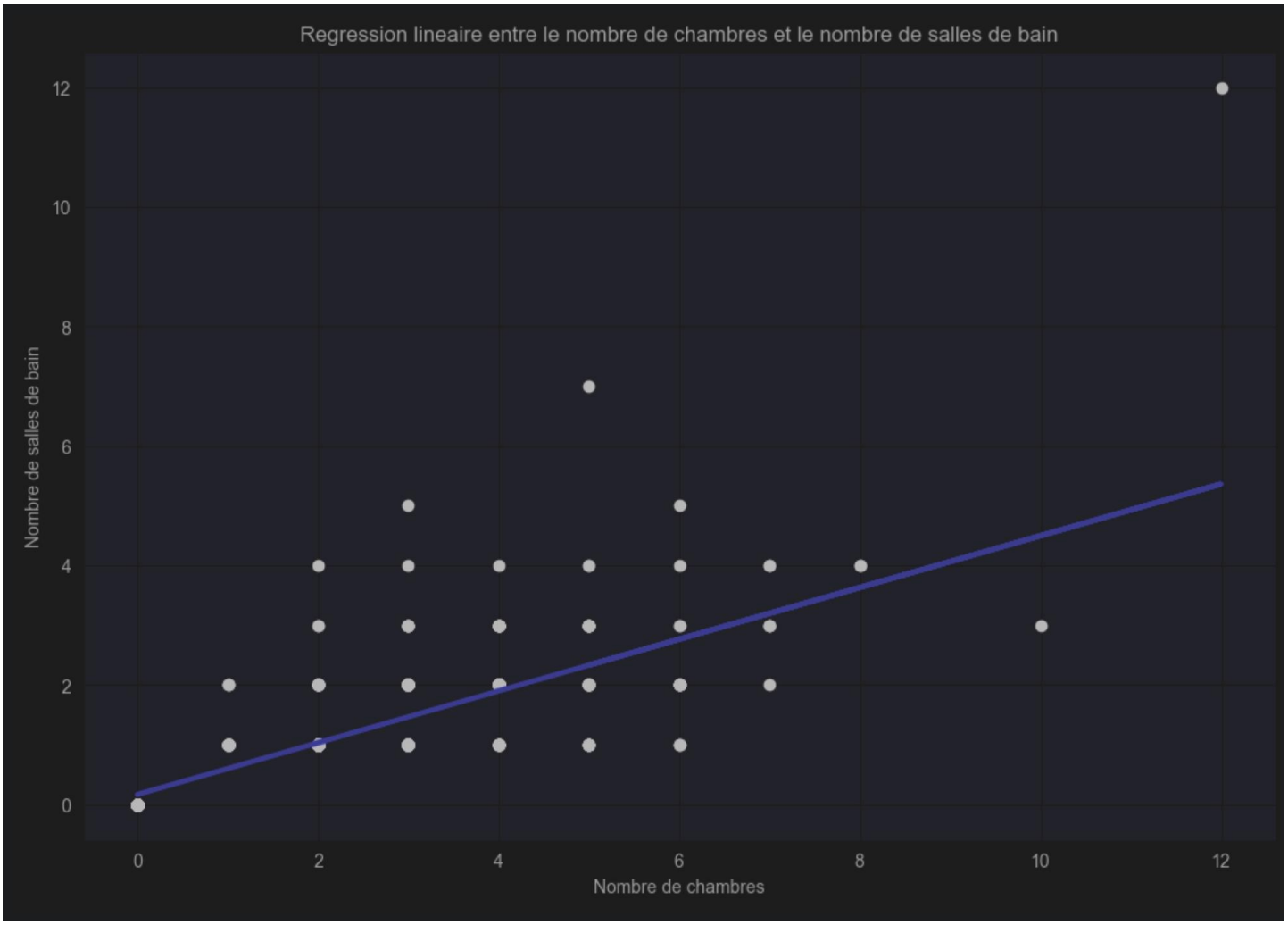
4. Algorithmes de régression

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.




```
Chambres        Salles de bain    0.83
                Assurances        0.71
Salles de bain  Chambres          0.83
Assurances      Chambres          0.71
dtype: float64
```

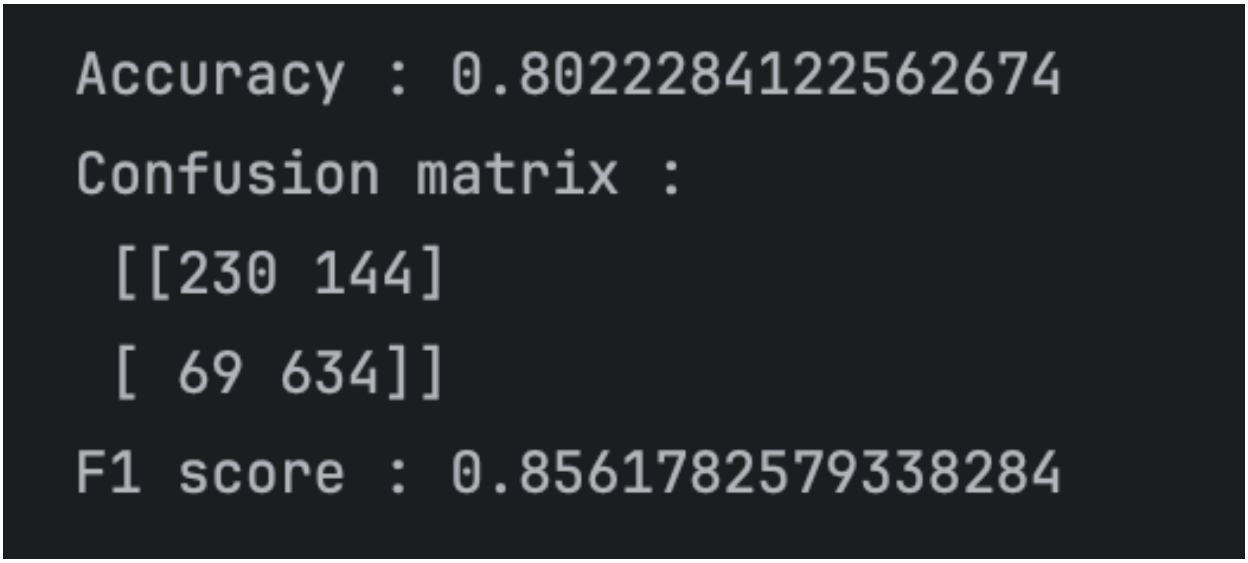
On a décidé de prendre la relation Chambre et Salle de bain pour établir un modèle de régression simple, avec un coefficient de corrélation de 0.83, ce qu'indiquent qu'il a une forte corrélation positive entre les deux.



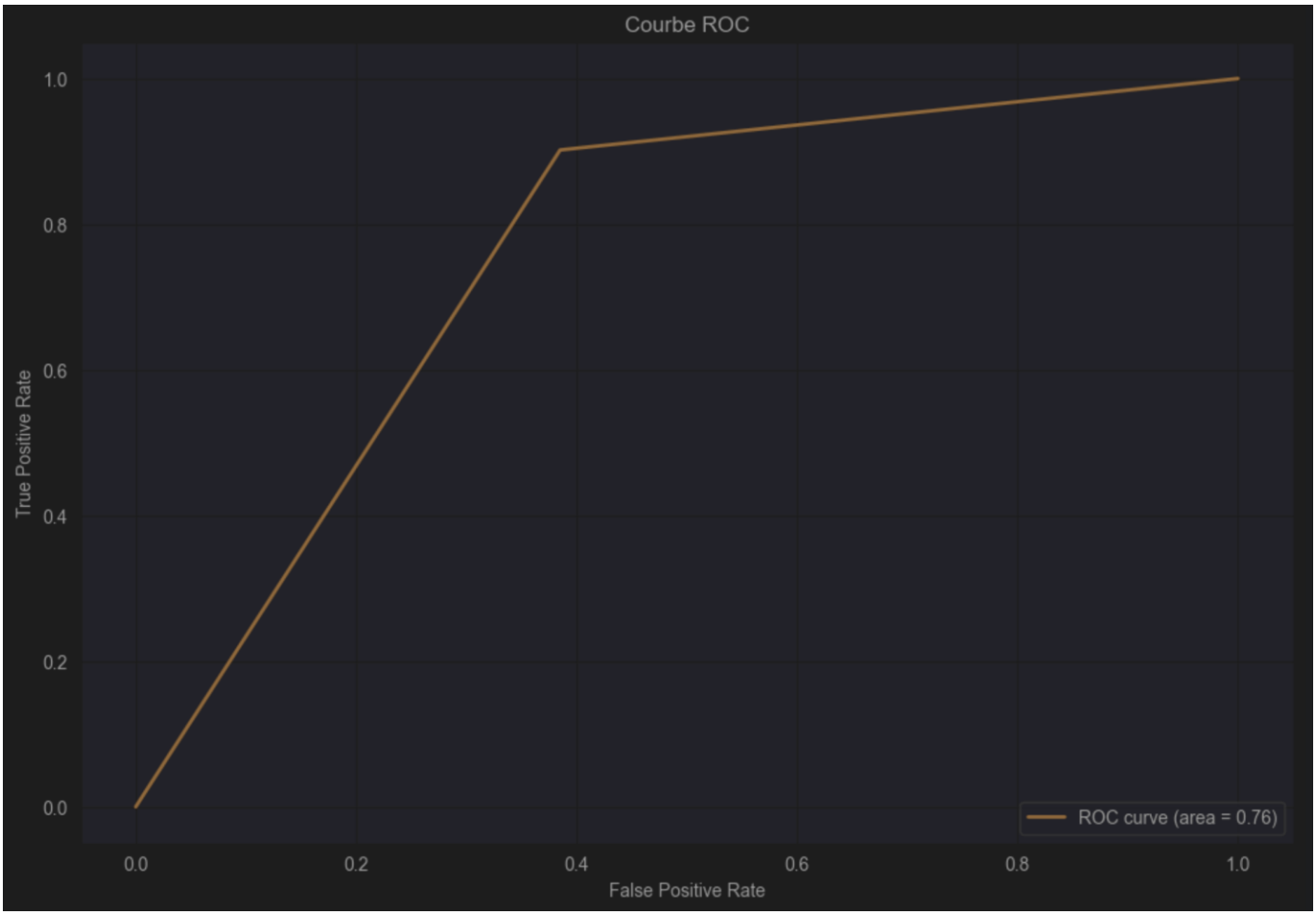

```
R2 : 0.6930308941720327
RMSE : 0.5568620602205876
R2 ajuste : 0.6927453415154485
```

Voici les résultats de notre modèle. On a un R2 de 0.69 ce qui signifie que notre modèle explique 69% de la variabilité des donnes de test, c'est un bon résultat dans cette situation de régression linéaire simple. On a aussi une erreur quadratique moyenne de 0.56 ce qui signifie que le modèle dans ces erreurs prédira une salle de bain de plus que ce qu'on attendait car nous sommes dans une prédiction d'entier dans le cas d'une salle de bain. Le R2ajuste est très proche du R2 normal car on est dans une régression linéaire simple, il y a qu'une seule variable explicative donc la pénalisation des variables est de fait minimal. On peut voir aussi sur le graphique que la régression suit la tendance des données. Je pense aussi que le point 10 chambres-3salle de bain est un point influent qui a influencé sur la pente de la régression.

2. Dans cette question, on s'int´eresse a` pr´edire si le prix d'une annonce sera sup´erieur ou inf´erieur a` 350000$ en fonction de la r´egion, du nombre de chambres, le nombre de salles de bain, le nombre de salles d'eau, le nombre d´´etages, la superficie de l'aire habitable, la taille du terrain, les taxes municipales et les taxes scolaires. Concevez un mod`ele de r´egression qui permet de faire cette pr´ediction et ´evaluer votre mod`ele.



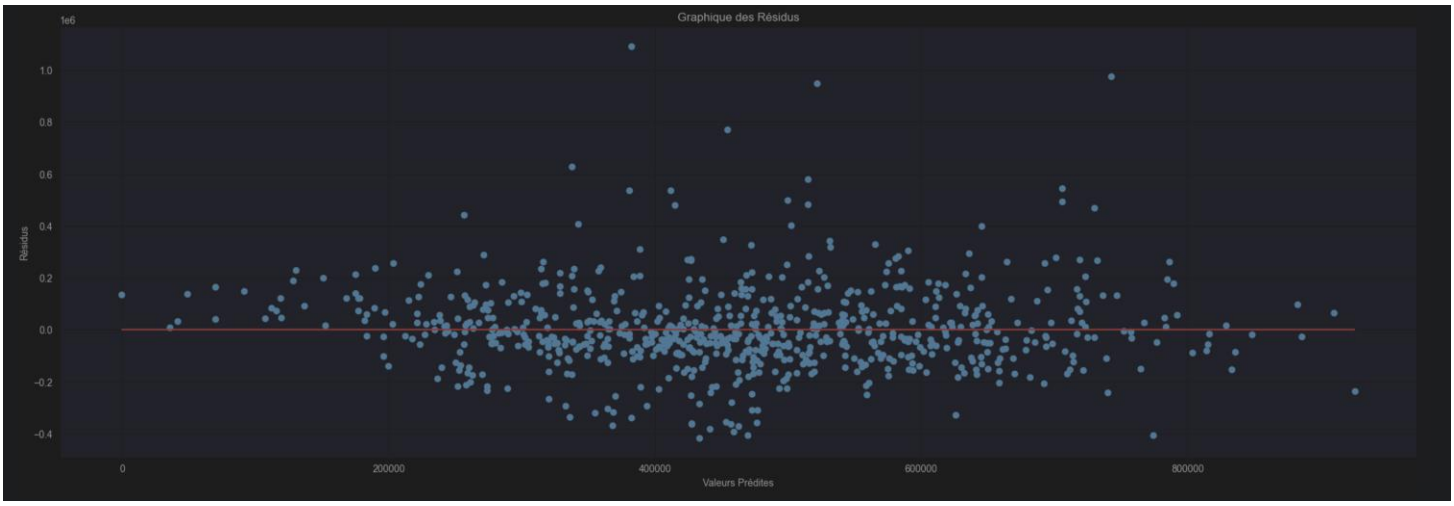
La précision de notre modèle de régression logistique est de 80% avec une score F1 de 0.86 ce qui est plutôt bon et que notre modèle a un certain équilibre entre la précision et le rappel.

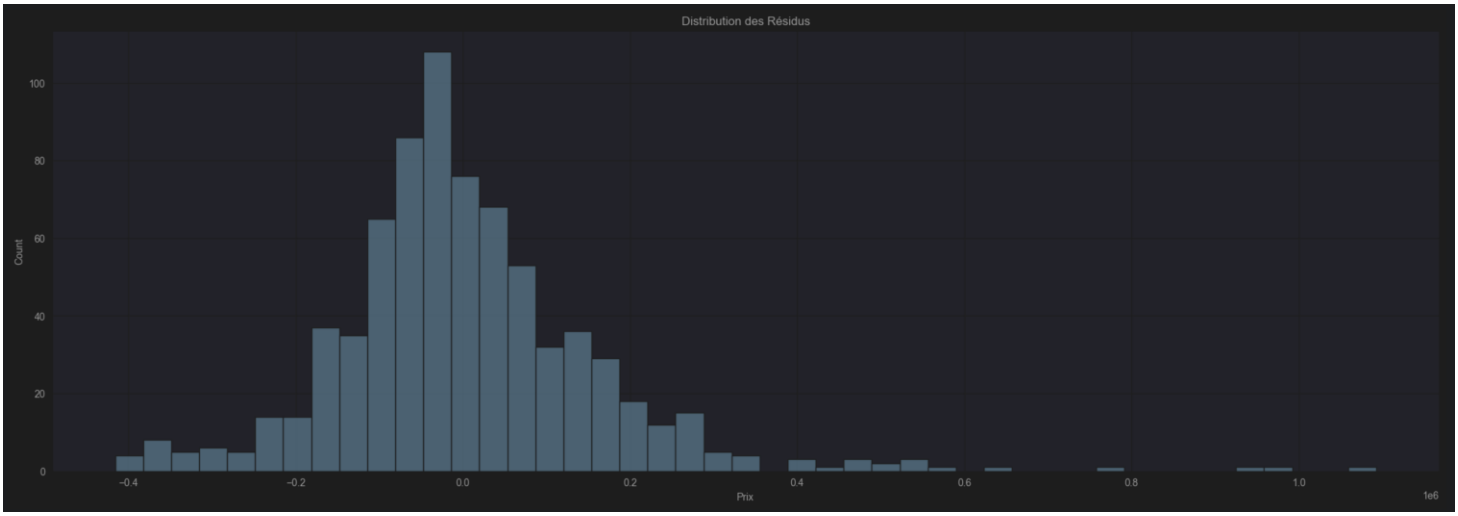


Pour des modèles de régression logistique on peut utiliser la courbe ROC pour l'évaluation, dans notre cas on a un AUC de 0.76, c'est correct comme valeur ce n'est pas nécessairement bon par contre on pourrait avoir mieux. Et la courbe montre une similitude d'une bonne courbe roc c'est dire monte haut vite puis devient plutôt stable.

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.

```
R2 : 0.4728501184576904
RMSE : 161918.91542501395
R2 ajuste : 0.45170579402514965
```





Pour cette question, j'ai pris la décision d'enlever les valeurs aberrantes qui avait vraiment un effet négatif sur la régression. Cela m'a permis d'avoir des prédictions plus robustes car les valeurs aberrantes peuvent fausser les résultats.

On obtient un R2 de 0.47 donc notre modèle explique 47% de la variation, on note une diminution légère de R2ajuster qui est à 0.45 cela veut dire que la pénalisation à affecter certaine variable explicative cela peut montrer un signe de multi colinéarité.

On observe une tendance vers l'homoscédasticité même si elle n'est pas parfaite. Sur la distribution on observer une tendance de loi normal en forme de cloche donc on peut dire qu'il n'y a pas d'erreur systématique dans les prédictions et qu'on reste proche du 0.

4. 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.

```
Propriété 1 : 651464.9249789726
Propriété 2 : 480160.92497897265
```

On obtient une évaluation de 651 464$ pour la première propriete et 480 160$ pour la seconde propriété.

5. Sans toutefois impl´ementer, pensez-vous que rajouter la ville dans vos 2 derniers mod`eles de r´egression con¸cue am´eliorerait la pr´ediction ? Justifiez votre r´eponse (un graphique ou un calcul).

Pour ma justification j'ai fait une analyse de variance ANOVA:

	sum_sq	df	F	PR(>F)
Ville	9.598214e+13	779.0	2.557173	1.410049e-79
Residual	2.041991e+14	4238.0	NaN	NaN

L'analyse de variance ANOVA nous indique que la variable Ville a un effet significatif sur le prix, ce qui est logique dans le contexte immobilier comme pour les régions sauf que les villes ont un rôle de plus grande précision sur les valeurs foncières.

Cependant dans notre cas nous avons 976 villes, ce qui est beaucoup trop pour notre modèle. Il faudrait donc faire une sélection de villes pour avoir un nombre raisonnable de villes. Car sinon nous pouvons exposer notre modèle a un surapprentissage. L'encodage de chez ville mettrait beaucoup de 0 aussi dans les colonnes encodées, ce qui pourrait aussi affecter la précision du modèle. Dans ce cas nous nous retrouverons avec trop de variable et ça serait difficilement interprétable. Donc dans notre cas, il serait préférable de ne pas ajouter la ville dans le modèle. Comme solution par exemple nous pourrions faire un modèle par région.

Partie 2:

1. Collecte de données

Comme dit en cours ils nous étaient impossible de ratisser le site de IMDB avec BS4, possible de le faire avec Sélénium par contre.

J'ai un code qui récupère sur les 50 premiers films afficher, cependant ils étaient impossibles de récupérer les genres, ils sont injectés via du JS après le render donc BS4 ne les capture avec une requête via requests.

Films: 0%| | 0/50 [00:00<?, ?film/s]

5 rows × 7 columns pd.DataFrame

	id	title	duration	gender_list	release_date	user_rating	nb_user_rating
0	4633694	Spider-Man: Into the Spider-Verse	1hour57minutes	None	December 14, 2018 (Canada)	8.4	641K
1	7784604	Hereditary	2hours7minutes	None	June 8, 2018 (Canada)	7.3	362K
2	5814060	The Nun	1hour36minutes	None	September 7, 2018 (Canada)	5.3	165K
3	2709692	Dr. Seuss' the Grinch	1hour25minutes	None	November 9, 2018 (Canada)	6.4	81K
4	4154756	Avengers: Infinity War	2hours29minutes	None	April 27, 2018 (Canada)	8.4	1.2M

2. Exploration des données

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).

Voir notebooks de la partie 2.

Remarques :

- J'ai suivi les remarques non ordonnee
- Pour avoir la colonne action j'ai du prendre le top 6 et non le top 5 car il se trouve en 6eme position après mon nettoyage
- J'ai supprimé la colonne date de sortie et renommer la colonne date ext en date de sortie
- Après plusieurs essais d'imputation par la moyenne, la médiane et le mode pour le remplacement d'utilisateur note, les meilleurs résultats étaient avec la suppression des valeurs manquantes, le jeu de donnée était suffisant pour appliquer la suppression.

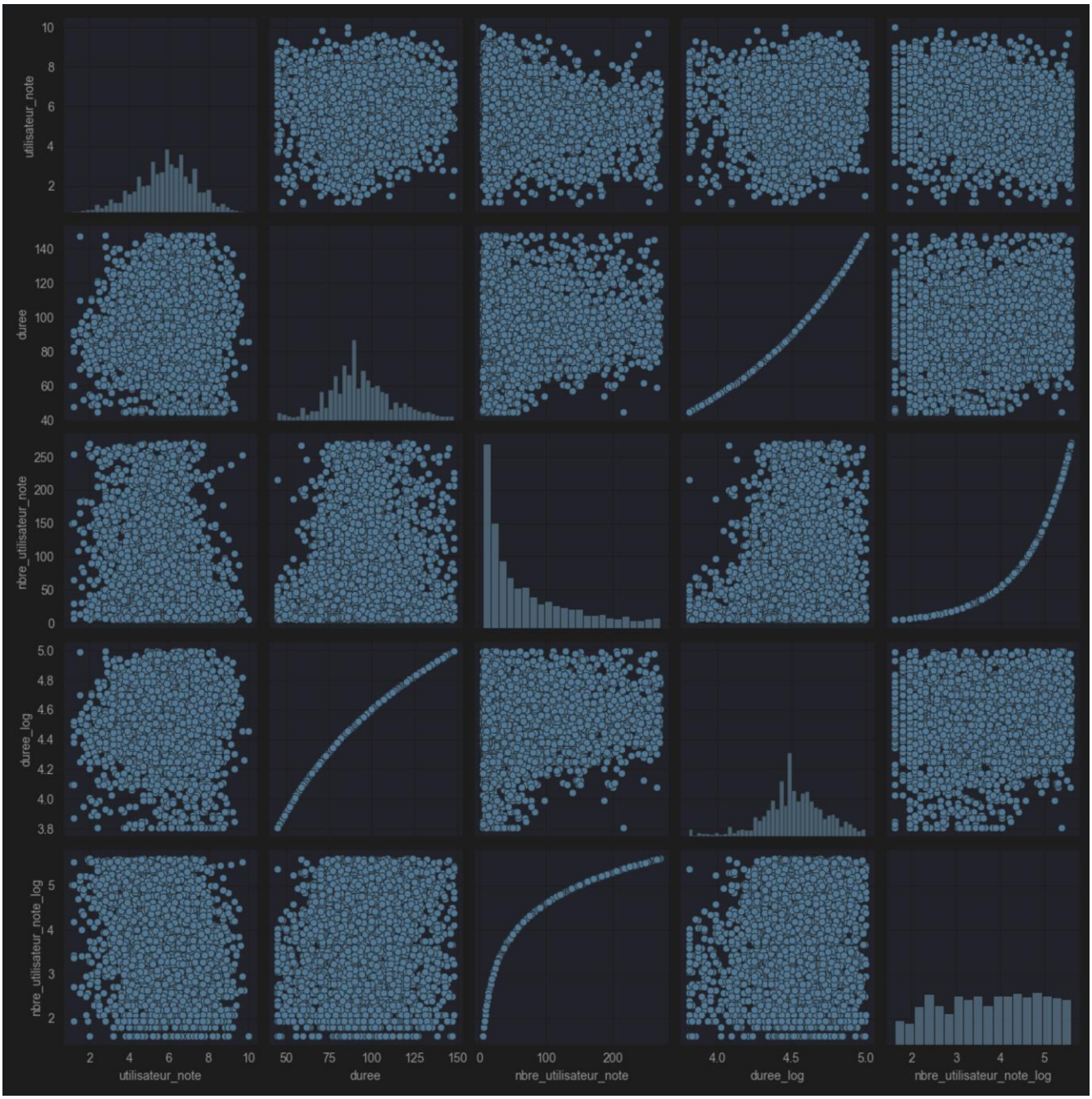
2. Créer 2 nouvelles colonnes durée minutes log, nbre utilisateur note log.

1-10 6807 rows × 16 columns pd.DataFrame CSV

	titre_film	utilisateur_note	genres_liste	duree	id_film	nbre_utilisateur_note	date_sortie	Drama	Comedy	Thriller	Horror	Romance	Action	autres_genres	duree_log	nbre_utilisateur_note_log
258	Virndavani Vairagya	6.6	[Drama]	91	tt9475282	25	2018-04-03	1	0	0	0	0	0	0	4.510860	3.218876
280	Peripheral	4.8	[Horror, Sci-Fi]	89	tt5658672	224	2020-08-03	0	0	0	1	0	0	1	4.488636	5.411646
297	Mar	5.5	[Drama]	103	tt7490386	49	2019-05-16	1	0	0	0	0	0	0	4.634729	3.891820
318	Exploitation	4.1	[Comedy]	111	tt7543138	61	2018-04-24	0	1	0	0	0	0	0	4.709530	4.110874
364	Can't Have You	3.7	[Comedy, Drama, Romance]	90	tt7141148	145	2018-03-07	1	1	0	0	1	0	0	4.499810	4.976734
447	His Perfect Obsession	4.9	[Thriller]	85	tt8105846	251	2020-10-26	0	0	1	0	0	0	0	4.442651	5.525453
496	Acute Misfortune	6.6	[Biography, Drama]	91	tt8634406	249	2018-08-04	1	0	0	0	0	0	1	4.510860	5.517453
540	Yinz	5.1	[Thriller]	89	tt5233090	120	2018-10-19	0	0	1	0	0	0	0	4.488636	4.787492
579	Enthusiastic Sinners	6.1	[Drama, Romance]	85	tt6888362	131	2019-10-08	1	0	0	0	1	0	0	4.442651	4.875197
664	My Wife's Secret	5.7	[Drama]	52	tt9503774	21	2018-12-27	1	0	0	0	0	0	0	3.951244	3.044522

J'ai applique la fonction log de numpy sur les colonnes en question en créant des nouvelles.

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

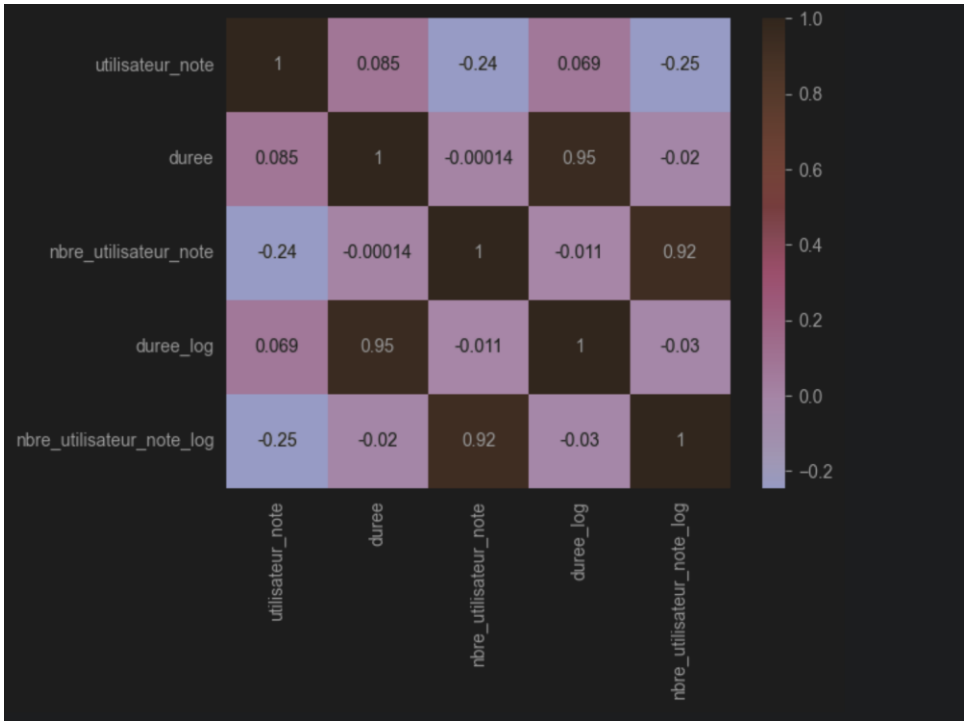


Pour faire l'analyse univarié comme pour la bivarié je vais utiliser ce pairplot. On peut utiliser la diagonale pour faire l'analyse univarié.

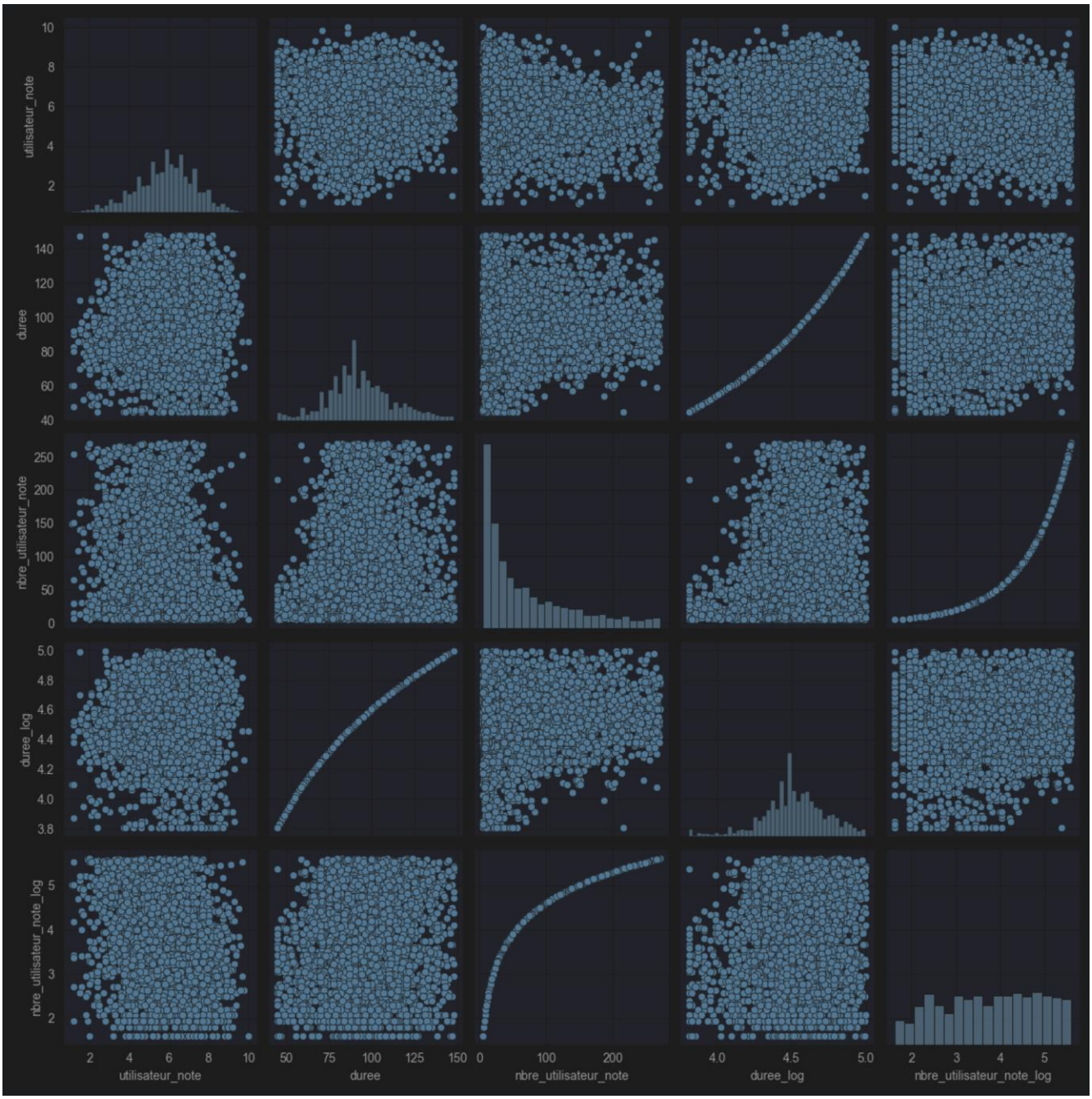
- Utilisateur_note: La distribution de utilisateur_note semble être assez uniforme avec un léger pic, ce qui nous porte à croire que la plupart des films ont une note similaire avec des légères variations, sachant que les notes vont de 1 à 10. Elle a l'air de suivre une loi gaussienne car elle est unimodale. Avec une légère asymétrie à droite.
- Duree: La distribution des durées des films semble unimodale, elle a l'air de suivre une loi gaussienne avec une légère asymétrie à gauche. Donc des films en moyenne de longueur normale en 75 et 100 minutes.

- Nbre_utilisateur_note: Cette variable semble être complètement asymétrique à gauche, on le voit à la longue que à droite et l'absence de queue à gauche.
- Duree_log: La transformation logarithmique de la durée semble avoir transférer l'asymétrie gauche à droite même si ça reste léger.
- Nbre_utilisateur_log: La transformation logarithmique a complètement enlevé l'asymétrie à gauche et cela a étalonner la distribution difficile de dire si la distribution est rendue normale.

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



D'après cette heatmap, on n'a pas de corrélation à part les colonnes log correspondante à leur origine ce qui est normale.

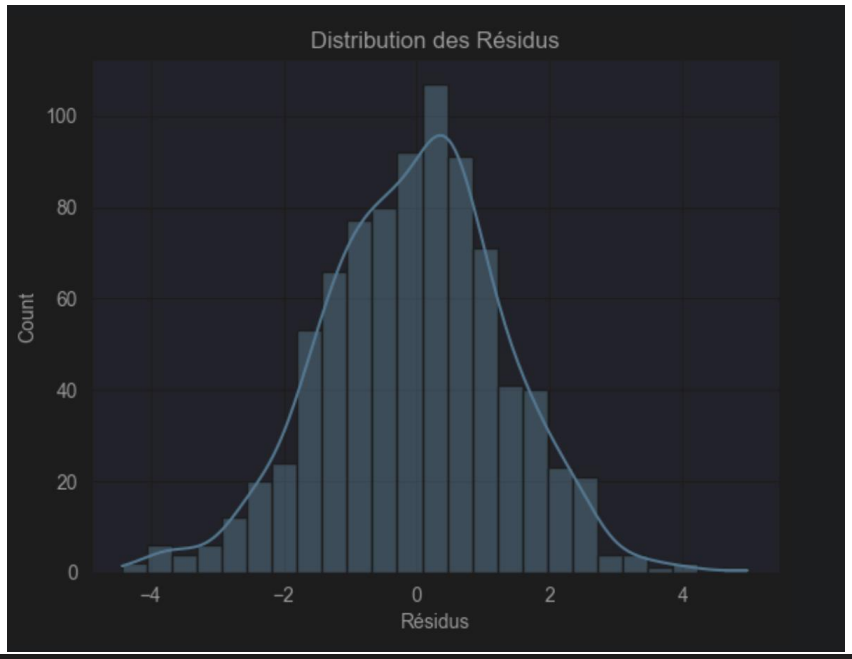


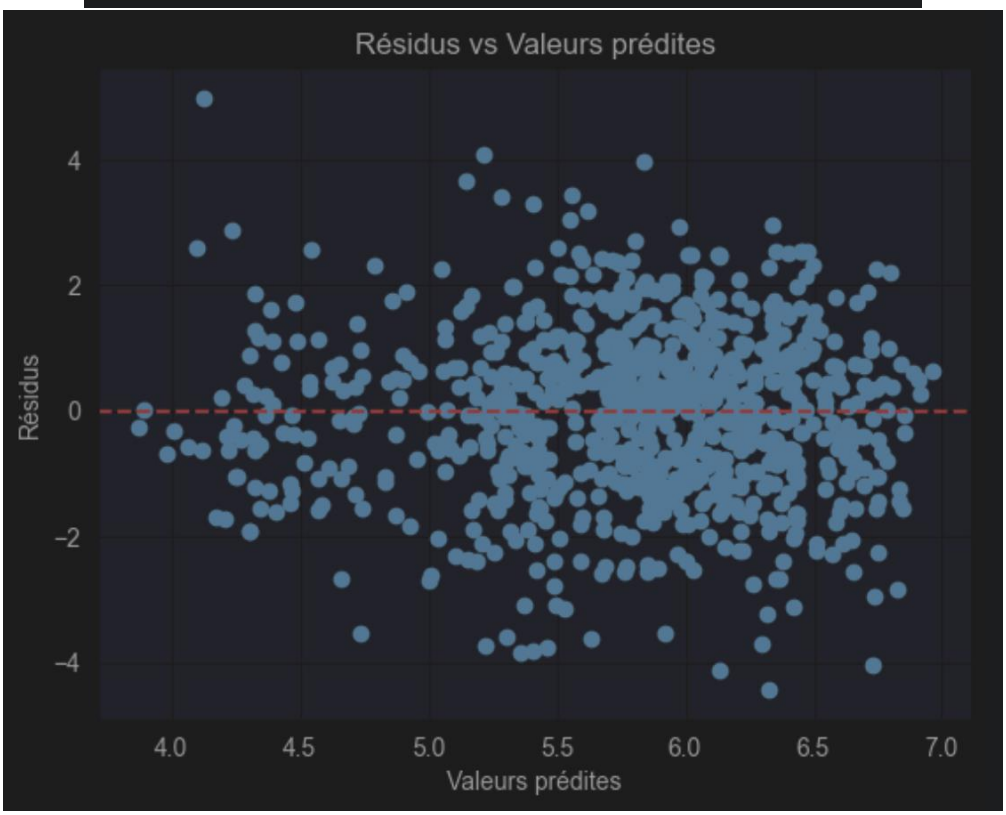
- Duree vs Utilisateur_note: On remarque une légère tendance ou quand le film est plus long la note est plus haute, on remarque ça car il y a un plus gros amas de point de donnée dans cette zone.
- Duree vs Nbr_utilisateur_vote: Cette relation montre un étalement très grand, mais on peut deviner une tendance où les films de plus longue durée ramassent plus de notes. On le voit à l'amas mais aussi à la courbe qui se créer quand le nombre de vote augmente.
- Note_utilisateur vs Nbr_utilisateur_vote: On observe ici une distribution qui suggère que les films avec des notes moyennes ont tendance à avoir un plus grand nombre de notes quand le nombre de votant augmente. On le voit comme un étau qui se resserre vers la moyenne quand le nombre de vote augmente.

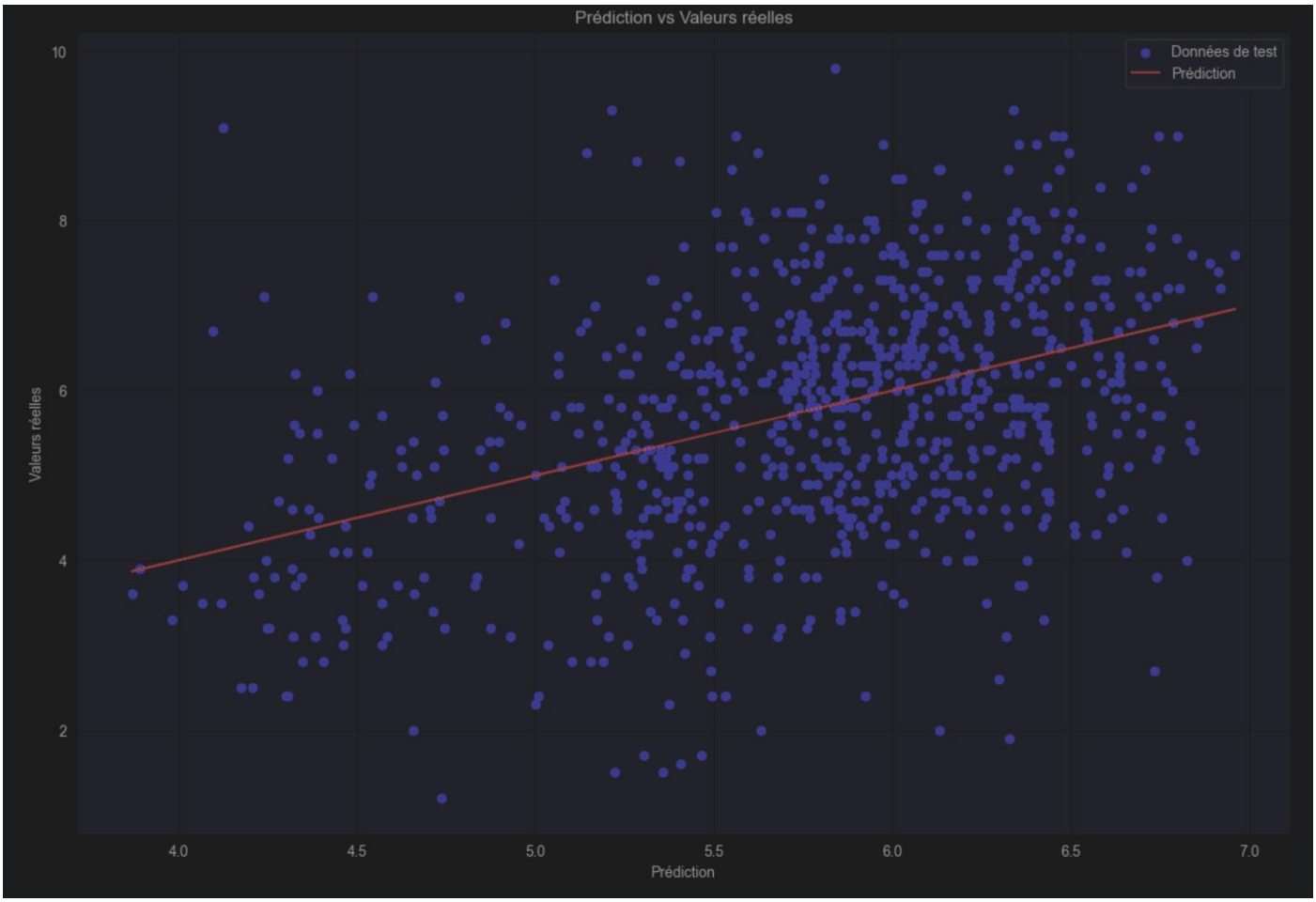
5. Dans cette question, on s'intéresse à prédire la note d'un film utilisateur note en fonction de 6 colonnes nbre utilisateur note log, durée minutes log, drame, action, thriller, et horreur.

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.

```
R2 : 0.171640144939877
MSE : 1.8296766456528166
r2adj : 0.16573032433302715
```



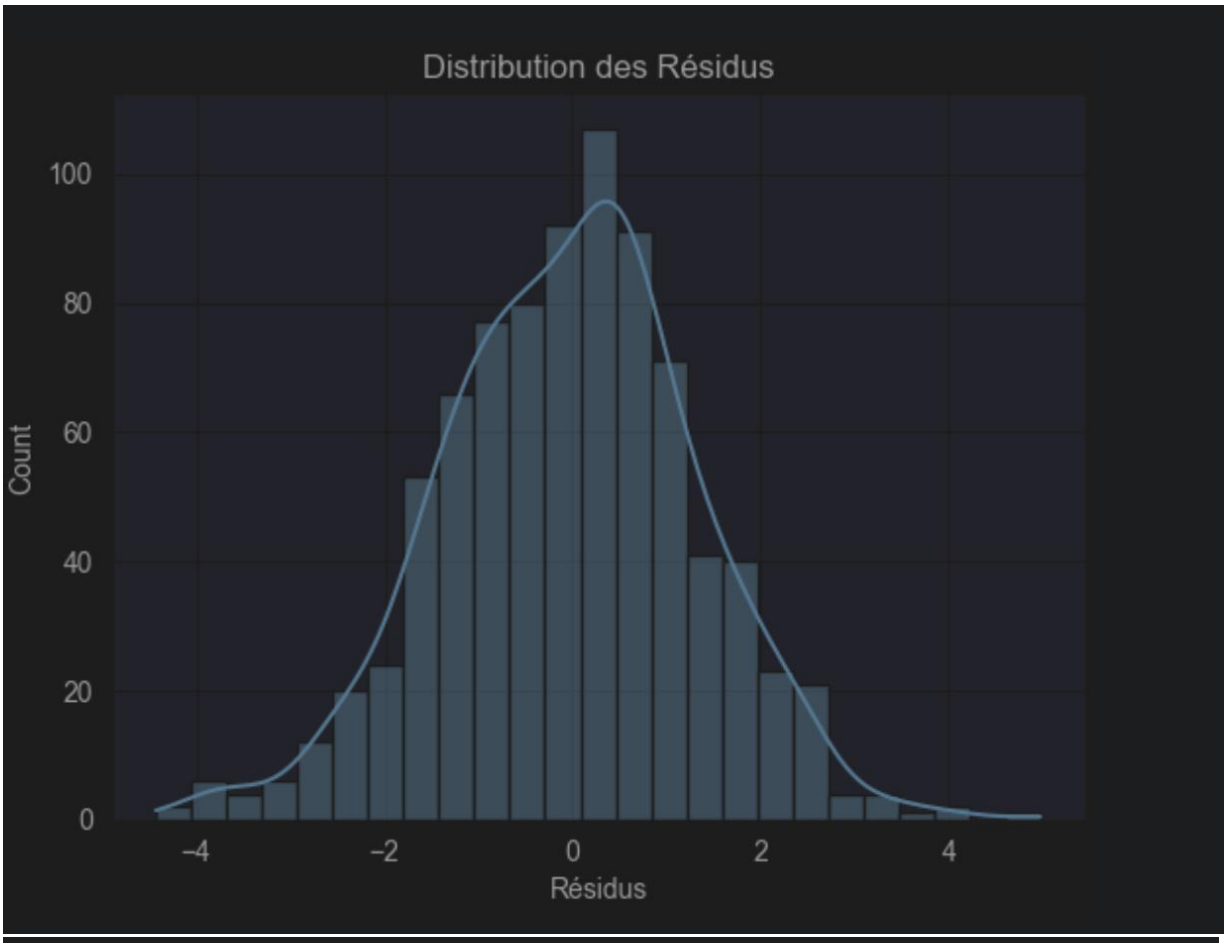


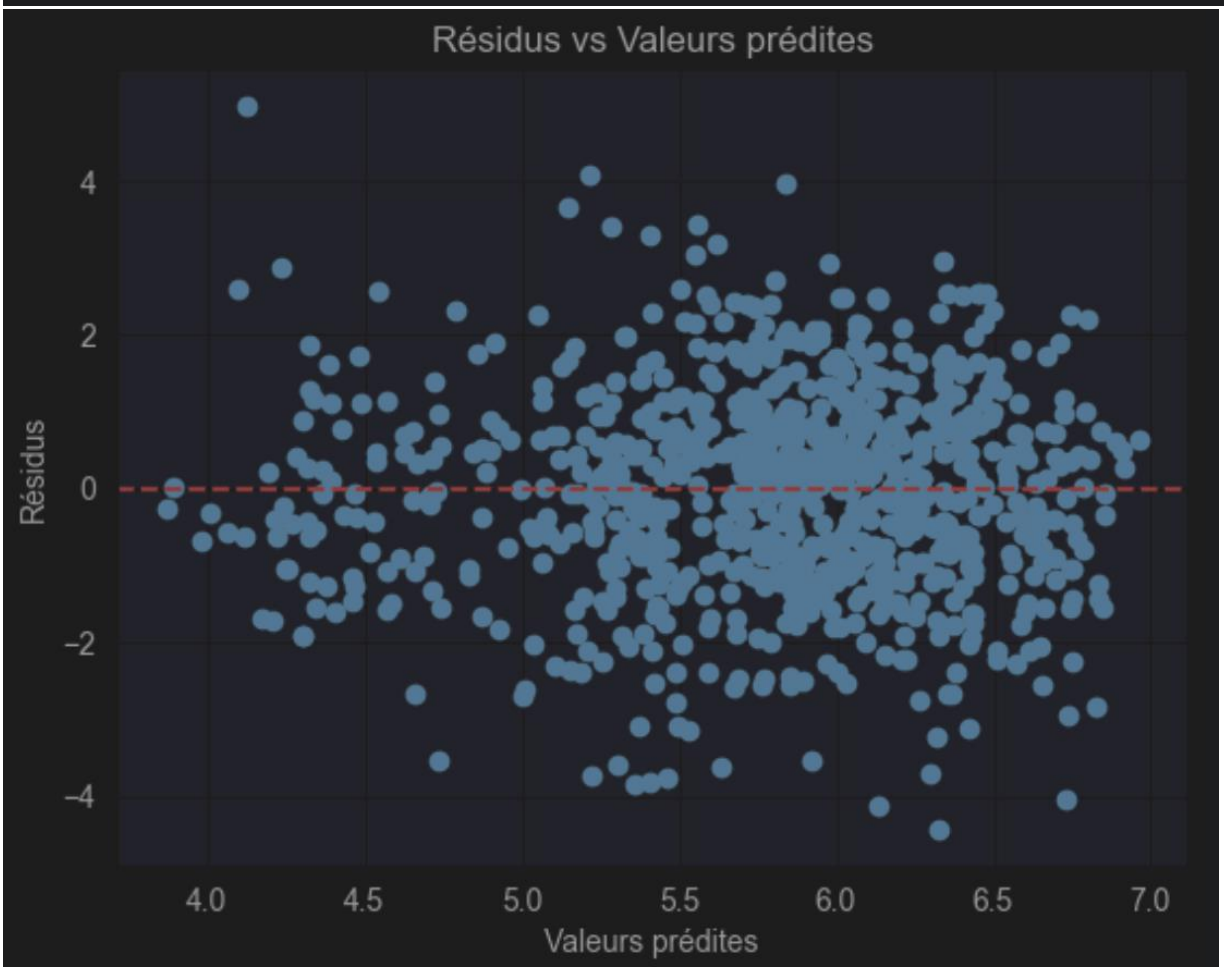


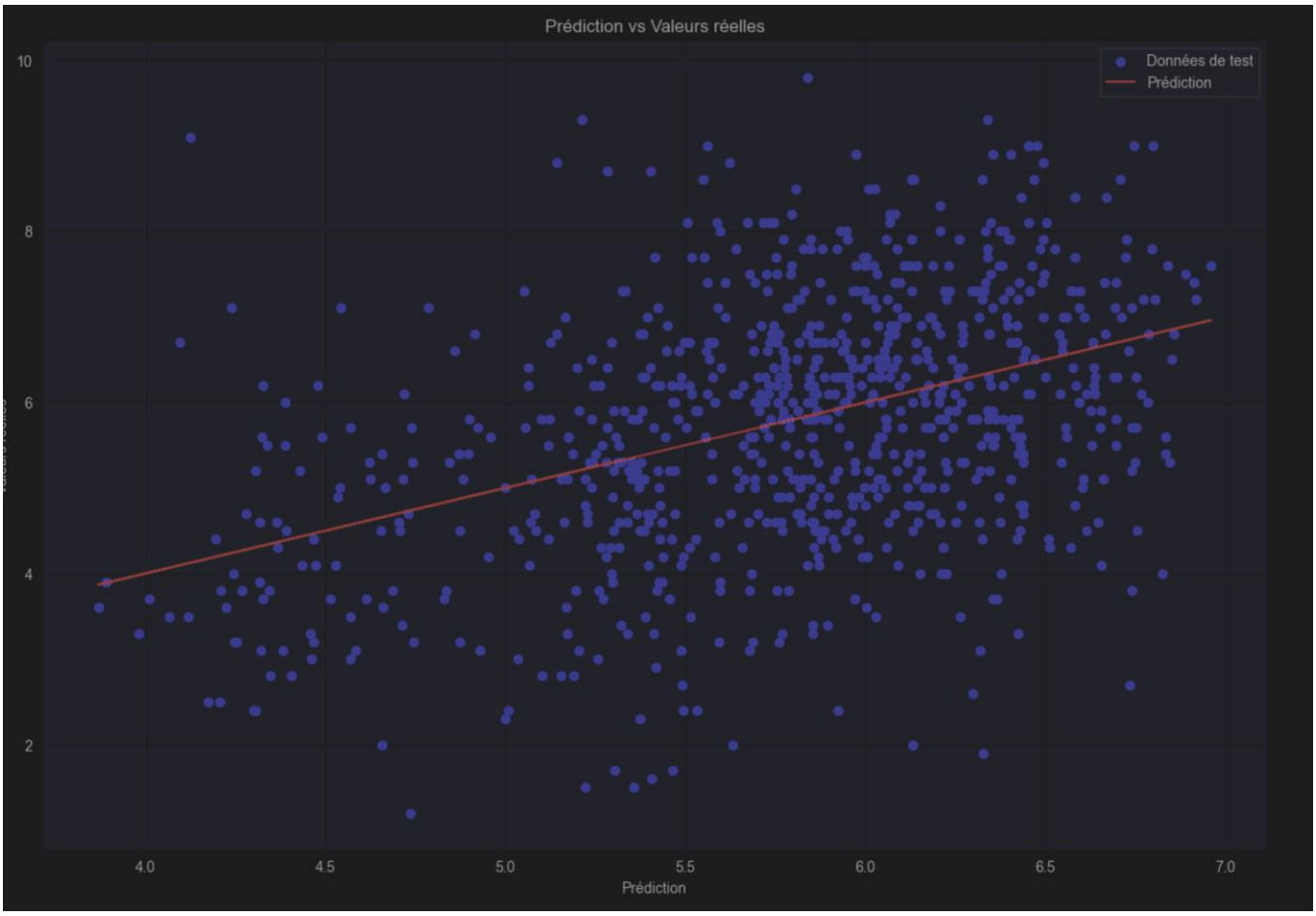
Pour cette régression on obtient des scores plutôt bas, avec un score R2 de 0.17, le modèle est capade d'expliquer seulement 17% des variations. Le R2 ajuste est de, 0.16, 16% ce qui fait qu'après pénalisation est moindre sur l'impact des valeurs explicatives. Dans l'analyses des résidus on est quand même proche du 0 et la distribution des résidus est en cloche donc le modèle s'ajuste quand même un peu. Pour essayer de comprendre plus j'ai fait le scatterplot des valeurs réelles en fonction des valeurs prédites. On remarque que c'est vraiment large autour de la droite de prédiction donc il est compréhensible que le modèle est du mal à expliquer la variance.

6. Refaire la question précédente après avoir effectué une normalisation adéquate. Comparer les résultats.

```
R2 : 0.171640144939877
MSE : 1.8296766456528166
r2adj : 0.16573032433302715
```


Distribution des Résidus
Count
Résidus
0
20
40
60
80
100
−4
−2
0
2
4


Résidus vs Valeurs prédites
Résidus
Valeurs prédites
4
2
0
−2
−4
4.0
4.5
5.0
5.5
6.0
6.5
7.0



On obtient des resultat identiques avec la normalisation. Mon hypothese c'est du fait qu'on a appliqué la fonction log sur les deux variables qui ont une grande échelle ce qui revient à une sorte de normalisation déjà faite lors du nettoyage. Car le fait d'appliquer le log on a déjà réduit l'asymétrie, stabiliser la variance et réduit l'échelle de grandeur. Et Pour les variables one hot encoder la-elle sont soit égale a 0 ou 1 donc la normalisation n'a pas d'incidence là-dessus.

Donc c'est normal d'obtenir les mêmes résultats après normalisation selon mon hypothèse.