République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université de Saad Dahleb Blida -1-
Faculté des Science
Département d'informatique

MASTER EN INGÉNIERIE DES SYSTÈMES INTELLIGENTES

ANALYSE DE DONNÉES

ÉTUDIENTS :
01. ABDELATIF MEKRI
02. HALIMA NFIDSA
03. NAHLA YASMINE MIHOUBI

SUPERVISÉ PAR :
Dr. M.FAREH

Année universitaire :2023/2024

TABLE DE MATIERES

Résumé :

L'objectif principal de ce mini-projet est de réaliser une analyse exhaustive sur un jeu de données spécifique. Cette analyse vise à explorer, identifier et comprendre les relations, les modèles et les tendances présents dans l'ensemble de données. Par le biais d'une approche séquentielle et progressive, le projet cherche à extraire des informations précieuses concernant les variables étudiées, leurs interactions et leurs influences réciproques. Ceci permettra d'obtenir une compréhension approfondie des données, facilitant ainsi la prise de décisions éclairées. Pour interpréter les données de manière efficace, le projet suit une méthodologie systématique, incluant des étapes d'analyse univariée, bivariée et multivariée. Ces étapes permettent d'explorer et d'interpréter différentes facettes des données, commençant par des analyses simples pour progresser vers des examens plus complexes et détaillés. Le langage de programmation Python sera utilisé pour cette analyse.

1. Introduction :

1.1 Introduction:

L'analyse de données est un voyage méthodique vers la compréhension. Ce projet se penche sur cette exploration en suivant une approche en trois étapes.

Premièrement, l'analyse univariée se concentre sur chaque variable indépendamment, fournissant des informations spécifiques à chacune.

Ensuite, l'analyse bivariée explore les relations entre deux variables, mettant en lumière les corrélations et les interactions directes.

Enfin, l'analyse multivariée plonge dans la complexité des relations entre plusieurs variables, révélant des modèles plus élaborés et des dépendances plus subtiles.

Chaque étape s'intègre dans une méthodologie rigoureuse pour dévoiler les secrets et les tendances enfouis dans notre ensemble de données.

1.2 Objectifs du Projet:

- Compréhension des Tendances des Prix des Maisons: Explorer et analyser les tendances historiques de l'Indice des Prix des Maisons (HPI) pour obtenir des insights sur l'évolution des prix des maisons individuelles au fil du temps. Identifier les périodes de croissance ou de déclin significatives des prix des maisons aux niveaux national et régional.
- Analyse Géographique: Mener une analyse géographique complète en examinant les fluctuations des prix des maisons à différents niveaux, y compris les divisions de recensement, les états, les zones métropolitaines, les comtés, les codes ZIP et les secteurs de recensement. Identifier les régions avec des variations notables dans les tendances des prix des maisons et comprendre les facteurs contribuant à ces variations.
- Compréhension de la Méthodologie: Acquérir une compréhension approfondie de la méthodologie du HPI, qui utilise une technique statistique pondérée de ventes répétées. Explorer comment le HPI intègre les données de dizaines de millions de ventes de maisons, fournissant une approche transparente et complète pour analyser les données de transactions de prix des maisons.
- Indicateurs Économiques: Évaluer le HPI en tant qu'indicateur économique, en comprenant sa pertinence pour estimer les changements dans les défauts de prêts

hypothécaires, les remboursements anticipés et l'accessibilité au logement. Enquêter sur la manière dont le HPI peut être utilisé comme un outil pour les économistes du logement afin d'améliorer les capacités analytiques dans des zones géographiques spécifiques.

- Exploration et Description de l'Ensemble de Données : Fournir une description détaillée de l'ensemble de données choisi, y compris le lien de téléchargement, le nombre d'attributs, le nombre d'observations (transactions de prix des maisons) et les types de variables (par exemple, géographiques, temporelles). Comprendre la structure et la composition de l'ensemble de données HPI pour assurer une analyse efficace.

1.3 Raisons du Choix de l'Ensemble de Données HPI :

- Couverture Complète : Données disponibles pour les 50 États et plus de 400 villes aux États-Unis, offrant une vision exhaustive des tendances des prix immobiliers à différents niveaux géographiques.

- Données Longitudinales : Disponibles depuis les années 1970, offrant une perspective historique pour analyser l'évolution des prix des maisons au fil du temps.

- Transparence et Méthodologie : Méthodologie transparente basée sur des techniques statistiques fiables, assurant une base solide pour l'analyse.

- Perspectives pour la Prise de Décision : Fournit des insights précieux pour les décideurs dans les secteurs immobiliers, financier et économique, avec des informations géographiques détaillées.

- Pertinence pour les Économistes du Logement : Outil analytique crucial pour estimer les tendances économiques régionales, comme les défauts de prêts hypothécaires ou l'accessibilité au logement.

2.Analyse Univariée :

Variables Étudiées : 'hpi_type', 'hpi_flavor', 'level', 'place_name', 'yr', 'index_nsa', et 'index_sa'

2.1 La variable 'hpi_type':

La variable 'hpi_type' présente quatre modalités distinctes : 'traditional', 'non-metro', 'distress-free' et 'developmental'. Sur un total de 123 970 enregistrements :

- La modalité la plus fréquente est 'traditional' avec 116 768 occurrences, représentant environ 94% de l'ensemble des données.
- Les autres modalités sont beaucoup moins fréquentes :
- 'non-metro' apparaît 5 405 fois (environ 4.36%)
- 'distress-free' a 1 572 occurrences (environ 1.27%)
- 'developmental' est la moins présente avec 225 occurrences (environ 0.18%).

La répartition en pourcentage confirme la prédominance significative de la modalité 'traditional', suivie par les autres modalités avec des fréquences beaucoup plus faibles.

Tableau statistique des modalités de 'hpi_type':

	Effictif	Fréquence	Effictif cumulé	Fréquence cumulée
traditional	116768	0.941905	116768	0.941905
non-metro	5405	0.043599	122173	0.985505
distress-free	1572	0.012680	123745	0.998185
developmental	225	0.001815	123970	1.000000

Figure 1:Tableau statistique des modalités de 'hpi_type'

Représentation graphique de 'hpi_type':

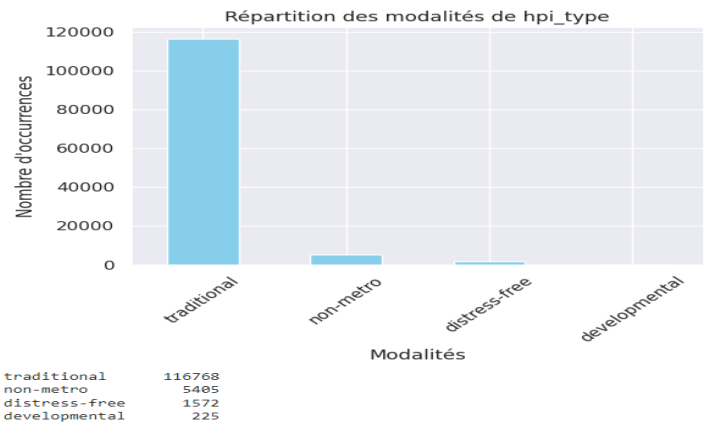

Figure 2:Diagramme de tuyaux de 'hpi_type'

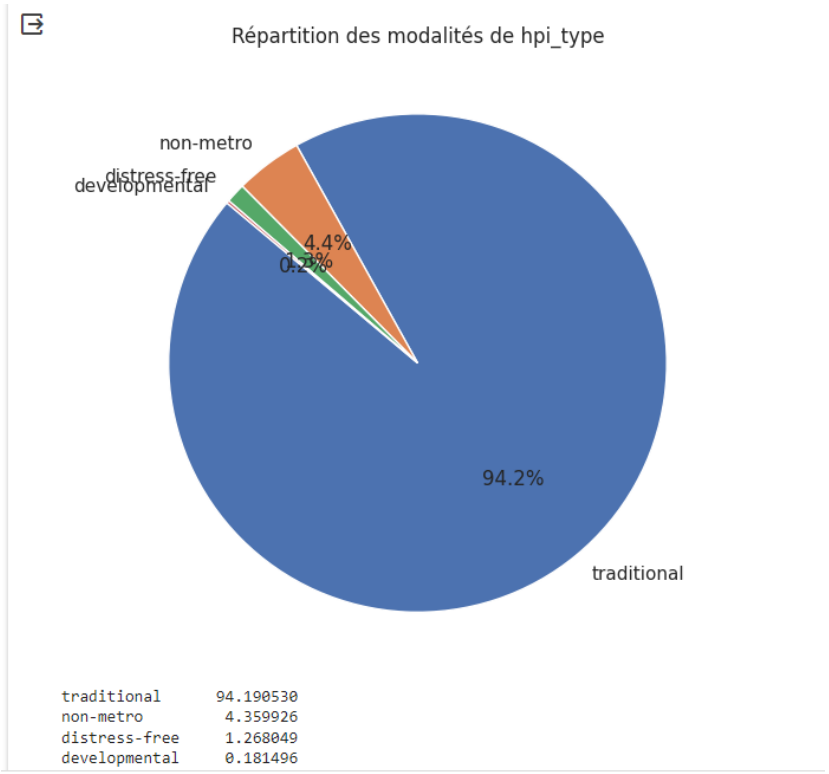

Figure 3:Diagramme circulaire de 'hpi_type'

2.2 La variable 'hpi_flavor':

La variable 'hpi_flavor' comporte trois modalités distinctes : 'purchase-only', 'all-transactions', et 'expanded-data'. Sur un total de 123 970 enregistrements :

- La modalité la plus fréquente est 'all-transactions', présente 82 725 fois, ce qui représente environ 66.73% de l'ensemble des données.
- Les autres modalités sont réparties comme suit :
    - 'purchase-only' apparaît 26 704 fois, soit environ 21.54%.

- 'expanded-data' est la moins fréquente avec 14 541 occurrences, représentant environ 11.73%.

En pourcentage, la modalité 'all-transactions' domine nettement, suivie par 'purchase-only' et 'expanded-data', qui ont des fréquences moins importantes.

Tableau statistique des modalités de 'hpi_flavor':

	Effictif	Fréquence	Effictif cumulé	Fréquence cumulée
all-transactions	82725	0.667299	82725	0.667299
purchase-only	26704	0.215407	109429	0.882705
expanded-data	14541	0.117295	123970	1.000000

Figure 4:Tableau statistique des modalités de 'hpi_flavor'

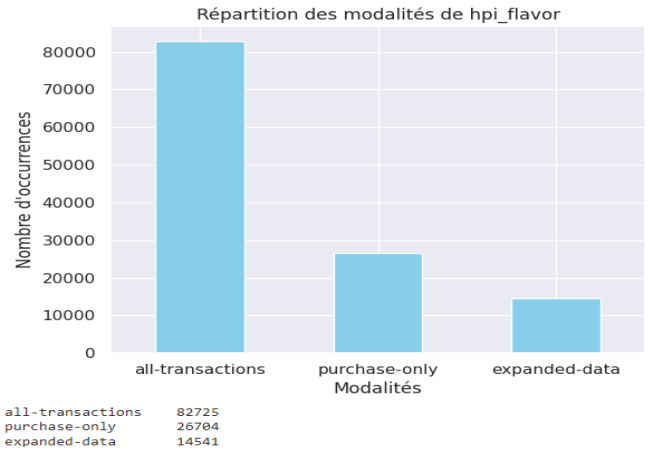

Figure 5:Diagramme de tuyaux de 'hpi_flavor'

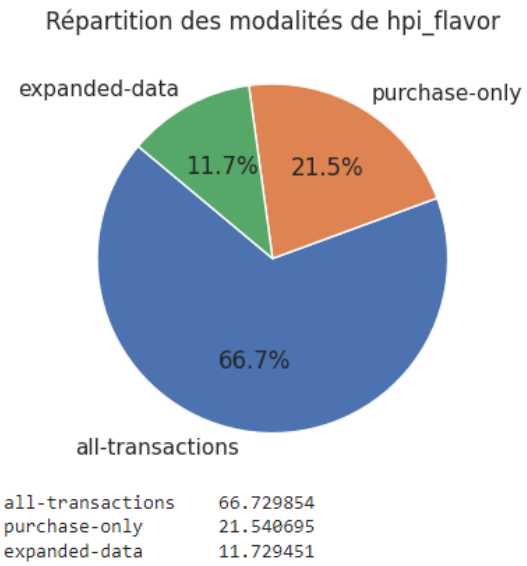

Figure6:Diagramme circulaire de 'hpi_flavor'

2.3 La variable 'place_name':

la variable 'place_name' comporte un total de 466 valeurs uniques avec une distribution inégale des fréquences. Voici un aperçu des 15 lieux les plus fréquents dans cette variable :

Les 10 premiers lieux ont tous une fréquence de 850, ce qui suggère une uniformité apparente dans ces catégories.

Ensuite, les données commencent à montrer des lieux spécifiques qui ne se répètent pas exactement 850 fois, mais qui sont parmi les plus fréquents.

Les quatre premiers lieux hors des 850 répétitions apparaissent environ 587 à 588 fois, suivis par d'autres lieux avec des fréquences similaires.

les lieux moins fréquents ont des occurrences aussi basses que 92, avec 'The Villages, FL' étant la moins fréquente.

Les 15 localités les plus fréquentes :

'East North Central Division', 'South Atlantic Division', 'East South Central Division', 'United States', 'West North Central Division', 'West South Central Division', 'Pacific Division', 'New England Division', 'Mountain Division', 'Middle Atlantic Division' sont les 15 localités les plus fréquentes, chacune apparaissant 850 fois.

Ces localités représentent les endroits les plus documentés dans cet ensemble de données, avec une fréquence égale pour les quinze premiers.

Les 15 localités les moins fréquentes :

'Pine Bluff, AR', 'Elizabethtown-Fort Knox, KY', 'Valdosta, GA', 'Weirton-Steubenville, WV-OH', 'Watertown-Fort Drum, NY', 'Sebring-Avon Park, FL', 'Bloomsburg-Berwick, PA', 'Gadsden, AL', 'Hammond, LA', 'Cumberland, MD-WV', 'Johnstown, PA', 'Beckley, WV', 'California-Lexington Park, MD', 'Hinesville, GA', 'The Villages, FL'.

Ces localités ont des occurrences plus faibles, allant de 122 à 92. Elles représentent les lieux moins documentés dans cet ensemble de données.

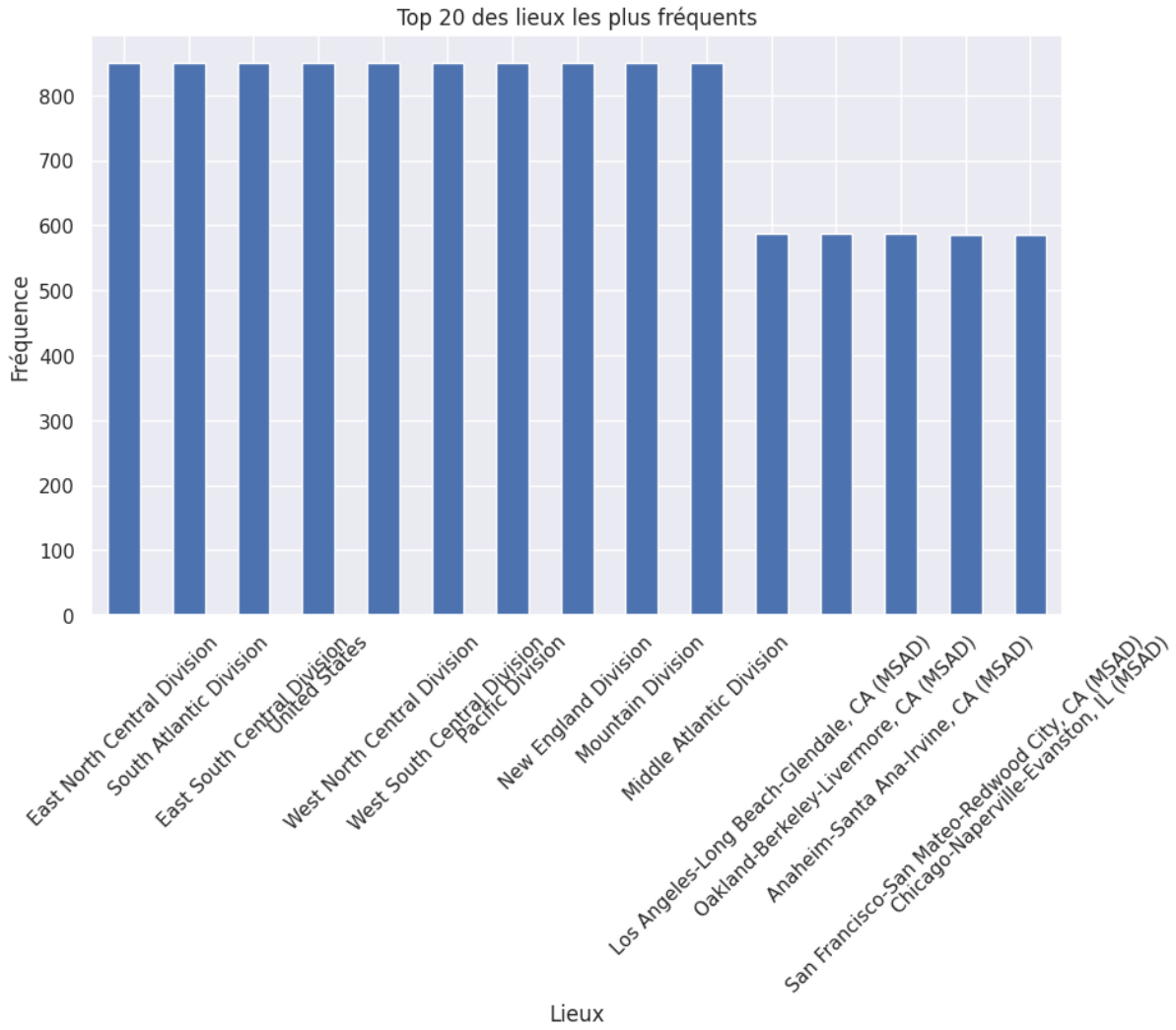

Figure 7:Diagramme de tuyaux de Les 15 localités les plus fréquentes

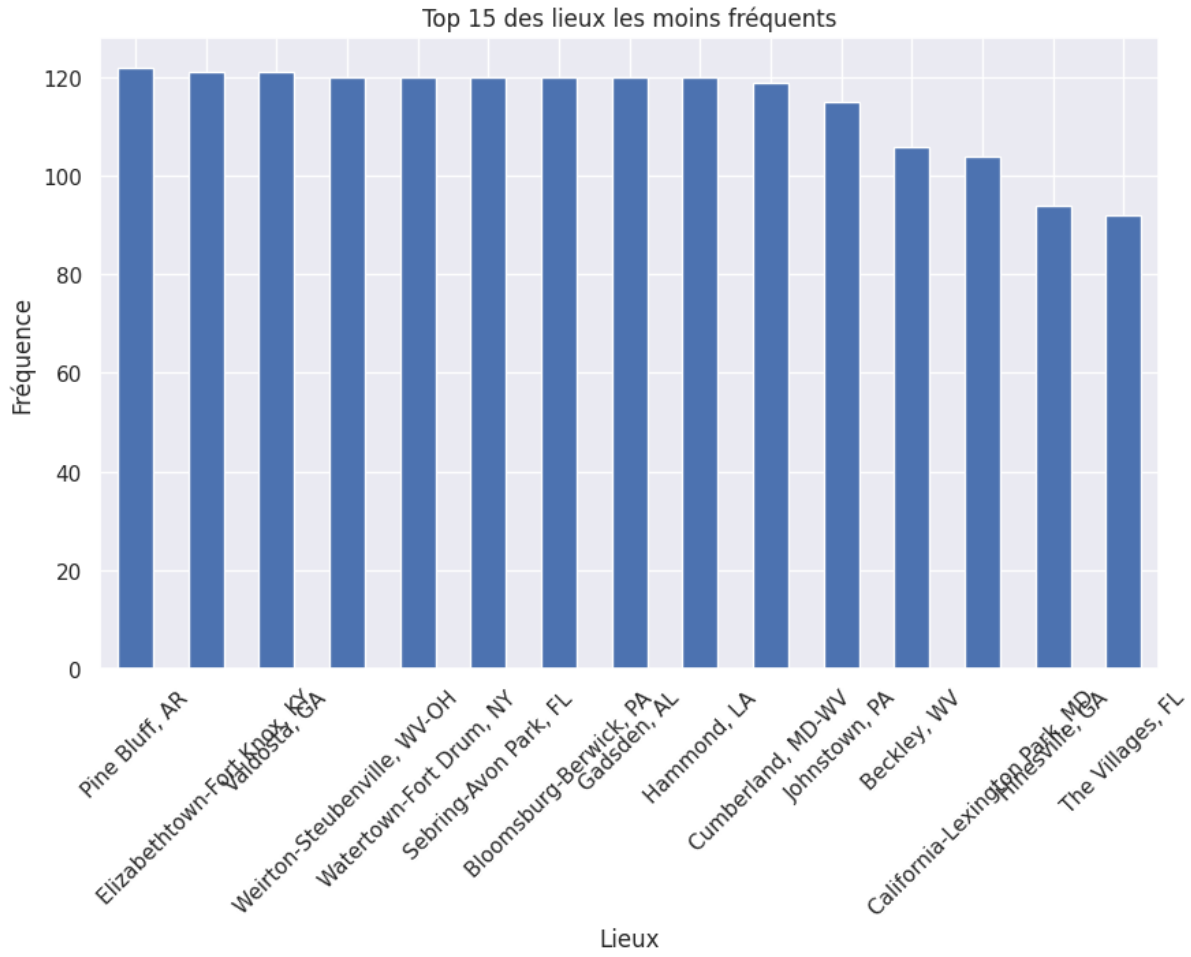

Figure 8:Diagramme de tuyaux de Les 15 localités les moins fréquentes

2.4 La variable 'level':

La variable 'level' répertorie quatre niveaux géographiques distincts, allant de niveaux plus larges tels que 'USA or Census Division' et 'State' à des niveaux plus spécifiques comme 'MSA' (Metropolitan Statistical Area) et 'Puerto Rico'. La catégorie la plus fréquente est 'MSA', apparue 86533 fois dans l'ensemble de données, suivie par 'State' avec 28712 occurrences.

'MSA' constitue la majorité des enregistrements avec une fréquence de 69.80%, indiquant une concentration significative des données sur les zones métropolitaines.

Ensuite, 'State' représente 23.16% des données, suivie par 'USA or Census Division' à 6.86% et 'Puerto Rico' à seulement 0.18%.

Cette distribution inégale des niveaux géographiques pourrait avoir un impact significatif sur les analyses régionales ou spatiales effectuées.

Tableau statistique des modalite de ‘level’:

	Effictif	Fréquence	Effictif cumulé	Fréquence cumulée
MSA	86533	0.698016	86533	0.698016
State	28712	0.231604	115245	0.929620
USA or Census Division	8500	0.068565	123745	0.998185
Puerto Rico	225	0.001815	123970	1.000000

Figure 9:Tableau statistique des modalité de 'level'

Représentation graphique:

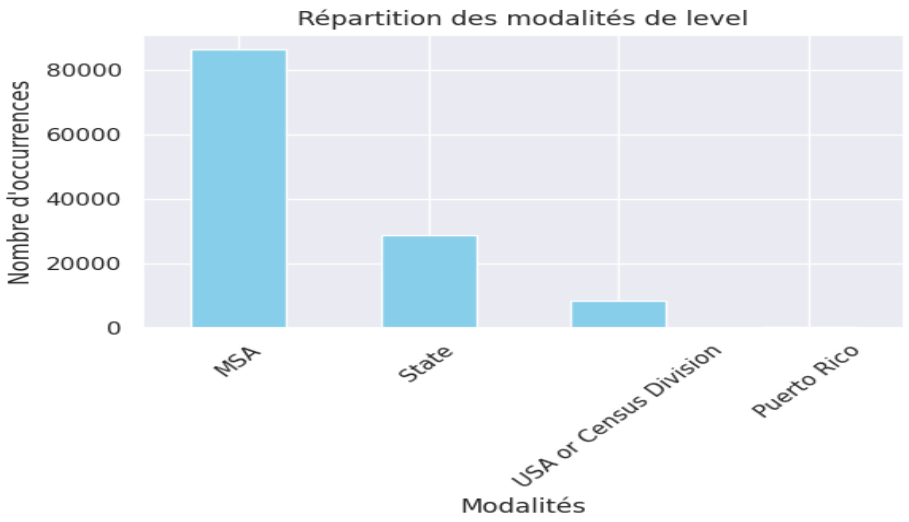

Figure10:diagramme de tuyaux de level

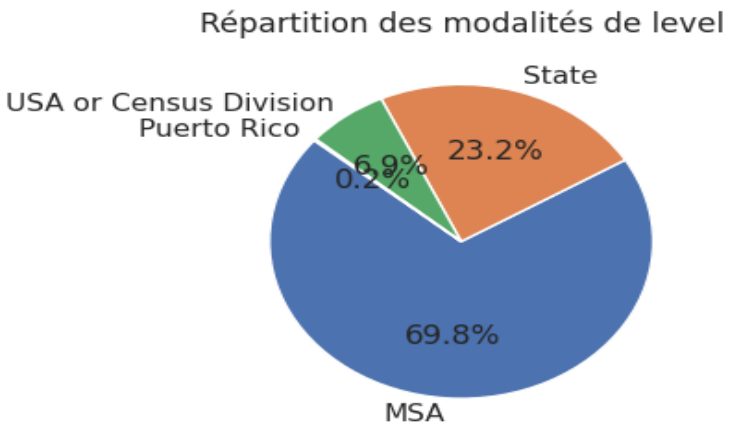

Figure11:diagramme circulair de level

2.5 La variable 'years':

La moyenne se situe autour de l'année 2004, illustrant le point central des données temporelles.

L'écart-type de ±11,76 années indique une dispersion autour de cette moyenne, soulignant une certaine variabilité dans les enregistrements.

Les quartiles indiquent que 50% des données se situent entre 1995 et 2014, avec des enregistrements allant de 1975 à 2023, soulignant une large période couverte par les données.

La médiane se trouve autour de 2005, montrant que la moitié des enregistrements sont avant cette année-là et l'autre moitié après.

De plus, l'analyse de l'asymétrie indique une légère asymétrie négative, ce qui suggère que la distribution des années a une queue gauche un peu plus longue que la droite par rapport à la moyenne.

L'aplatissement négatif montre également que la distribution est moins concentrée autour de la moyenne, avec des extrémités moins épaisses que celles d'une distribution normale, ce qui pourrait indiquer une dispersion plus large des années.

```
count    123970.000000
mean       2003.985634
std          11.761904
min        1975.000000
25%        1995.000000
50%        2005.000000
75%        2014.000000
max        2023.000000
```

```
Asymétrie: -0.2714510814807819
Aplatissement: -0.7968196927123388
```

Représentation graphique :

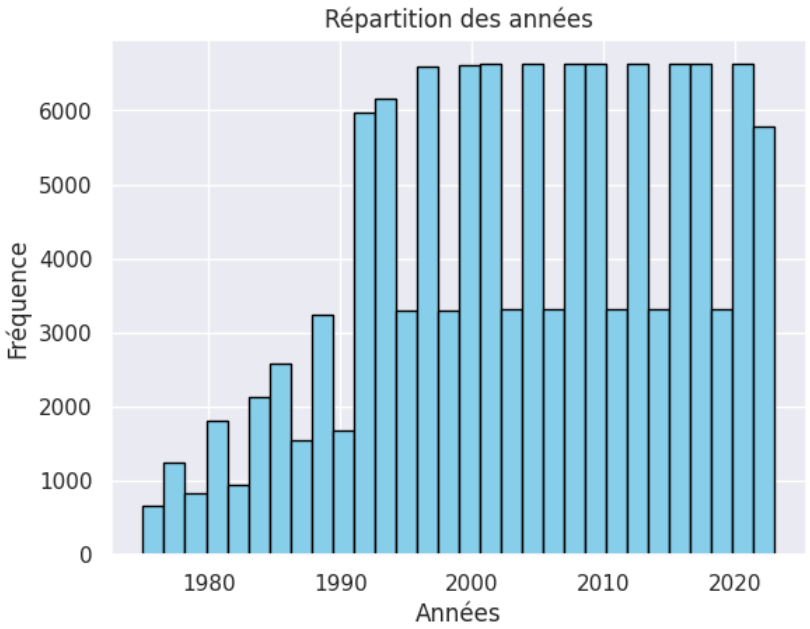

Figure12 : diagramme de tuyaux des années

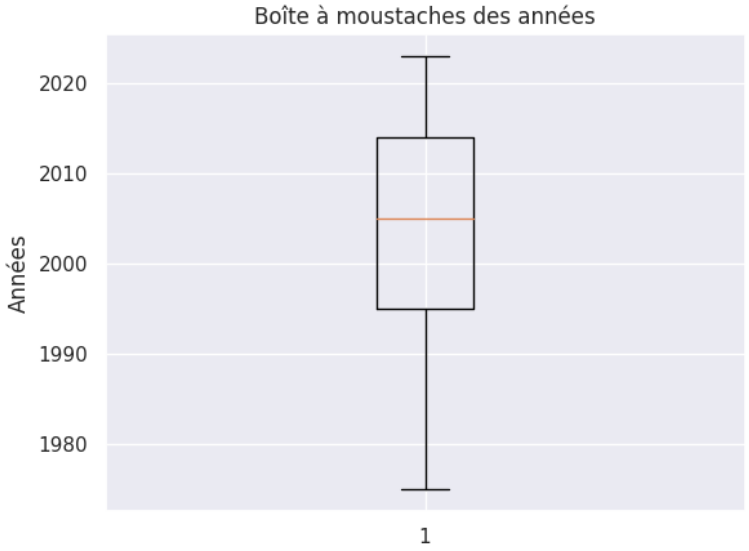

Figure 13: boite a moustaches des années

2.6 la variable 'Index NSA (Not Seasonally Adjusted)':

La moyenne se situe à environ 178.20, ce qui représente la tendance centrale des données.

L'écart-type de 97.41 indique une dispersion autour de cette moyenne.

Les valeurs s'étendent de 18.52 à 1181.99, montrant une grande variabilité dans les données.

Les quartiles (25%, 50%, 75%) situent la moitié des données entre 109.42 et 215.19, soulignant cette dispersion.

La variance élevée de 9488.65 et l'écart entre le minimum et le maximum de 1163.47 confirment cette variabilité importante.

L'asymétrie positive (2.03) indique une queue droite plus longue que la gauche par rapport à la moyenne, ce qui suggère une distribution où les valeurs extrêmes sont plus présentes du côté des valeurs élevées.

L'aplatissement élevé (7.45) indique que la distribution possède une concentration importante autour de la moyenne, avec des queues très épaisses, ce qui suggère également la présence de valeurs extrêmes éloignées de la moyenne.

```
count    123969.000000
mean        178.196744
std          97.409691
min          18.520000
25%         109.420000
50%         160.160000
75%         215.190000
max        1181.990000
Name: index_nsa, dtype: float64
Variance: 9488.647942328535
Étendue : 1163.47
Asymétrie: 2.0292903614350783
Aplatissement: 7.446864314027209
```

Représentation graphique :

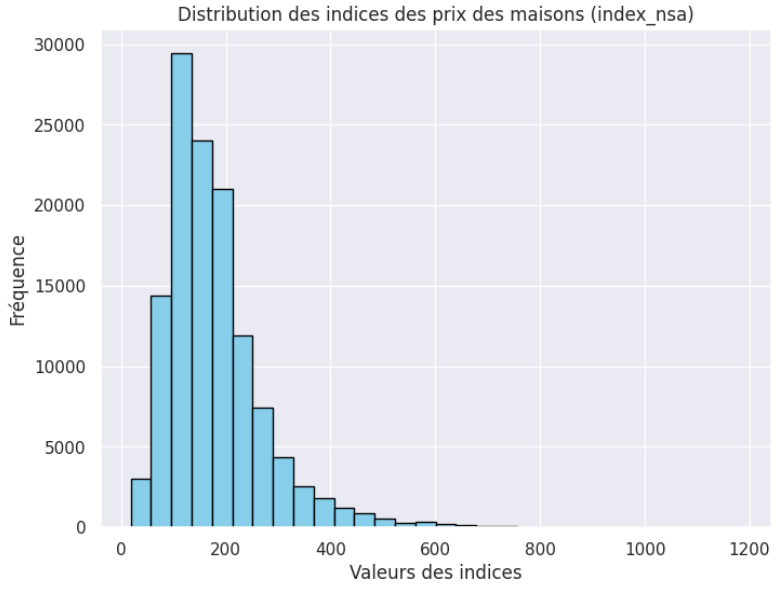

Figure 14:Distribution des indices des prix des maisons 'index_nsa'

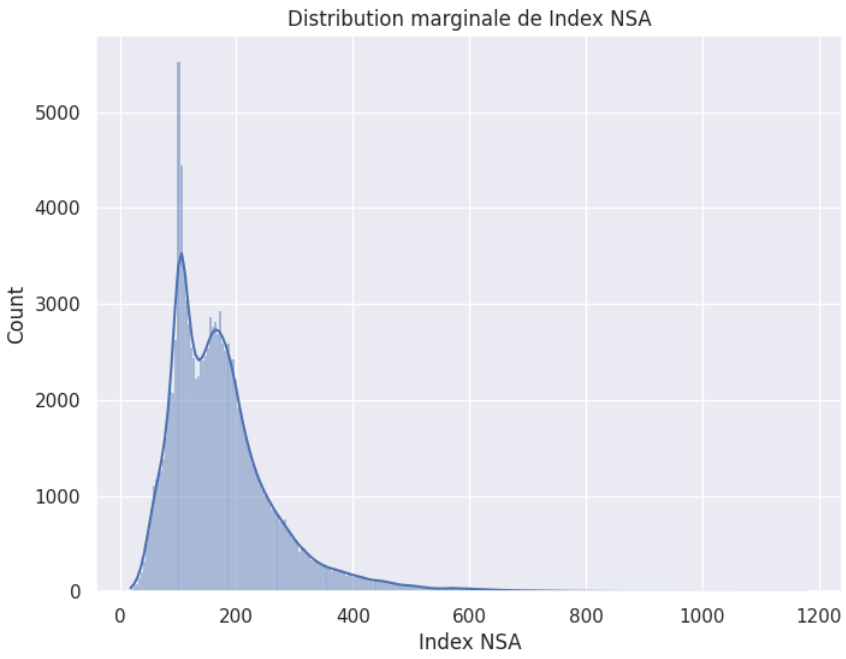

figure15 : Distribution marginale de 'indes_nsa'

2.7 la variable 'Index SA (Seasonally Adjusted)':

La variable 'Index SA (Seasonally Adjusted)' montre des caractéristiques distinctes :

Sa moyenne se situe autour de 199.67, représentant la tendance centrale des données.

L'écart-type de 92.21 indique une dispersion autour de cette moyenne.

Les valeurs varient de 74.79 à 828.16, illustrant une gamme considérable dans les données.

Les quartiles (25%, 50%, 75%) montrent que la moitié des données se trouvent entre 131.88 et 237.34, confirmant cette dispersion.

La variance de 8502.46 et l'écart entre le minimum et le maximum de 753.37 confirment la variabilité importante des données.

L'asymétrie positive (1.70) indique une queue droite légèrement plus longue que la gauche par rapport à la moyenne, suggérant une concentration plus importante vers les valeurs élevées.

L'aplatissement (4.42) montre que la distribution a une concentration autour de la moyenne, mais avec des queues moins épaisses que celles d'une distribution normale.

```
count    41245.000000
mean       199.669754
std         92.208774
min         74.790000
25%        131.880000
50%        183.390000
75%        237.340000
max        828.160000
Name: index_sa, dtype: float64
Variance: 8502.457925630328
Étendue : 753.37

Asymétrie: 1.699365566935375
Aplatissement: 4.423183370369203
```

Représentation graphique :

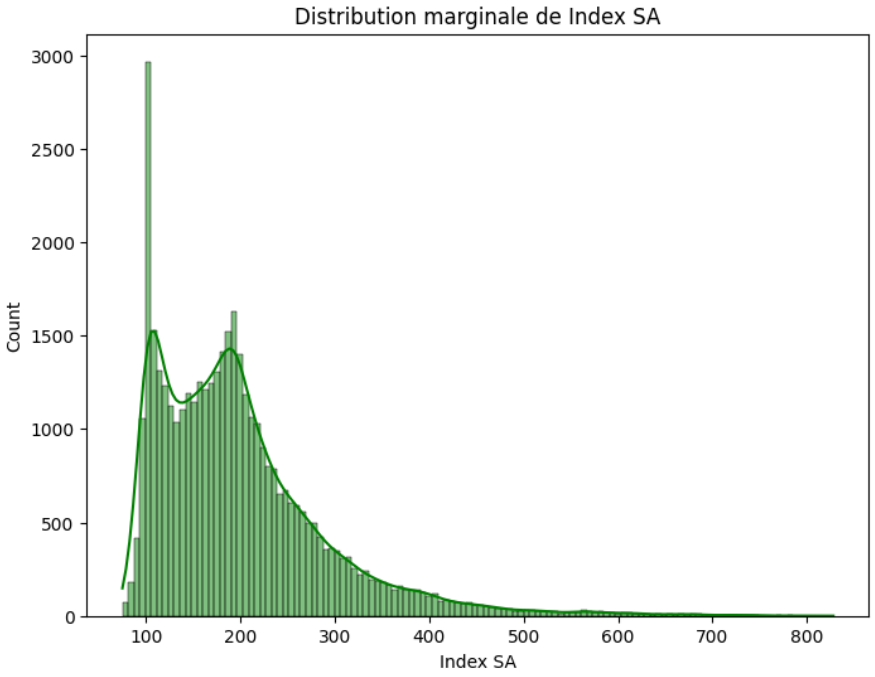

Figure 16: Distribution marginale de index_sa

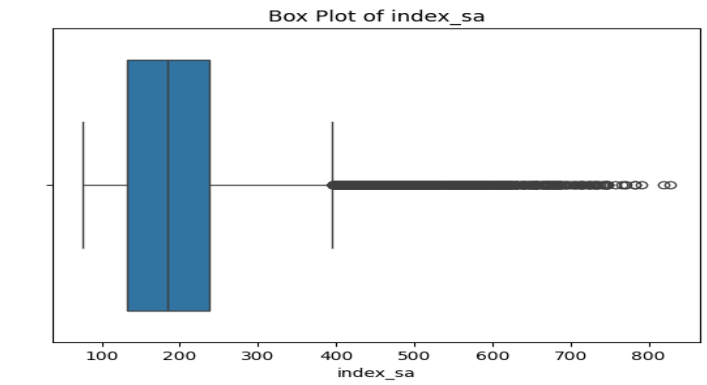

Figure 17: boite a moustaches de index_sa

3. Analyse Bivariée :

A.Variables Quantitatives:

A.1 L'étude entre 'index_sa' et 'index_nsa' :

Il existe une corrélation extrêmement forte entre les variables 'index_nsa' et 'index_sa'. La corrélation de 0,9997 indique une dépendance quasi totale entre ces deux variables. Lorsqu'une variable change, l'autre variable change proportionnellement, presque de manière identique.

La régression linéaire confirme cette forte corrélation avec un coefficient R proche de 1, ce qui suggère que les valeurs des indices SA et NSA varient ensemble selon une relation linéaire très étroite.

La représentation du nuage de points montre une dispersion des données le long d'une ligne droite, corroborant cette relation linéaire.

Il est important de noter que la suppression des valeurs manquantes a réduit considérablement la taille de notre ensemble de données, passant de 123 970 à 41 245 enregistrements. Cela indique qu'une partie significative de nos données avait des valeurs manquantes dans au moins l'une des variables 'index_nsa' ou 'index_sa'.

Matrice de corrélation :

	index_nsa	index_sa
index_nsa	1.000000	0.999739
index_sa	0.999739	1.000000

Représentation graphique :

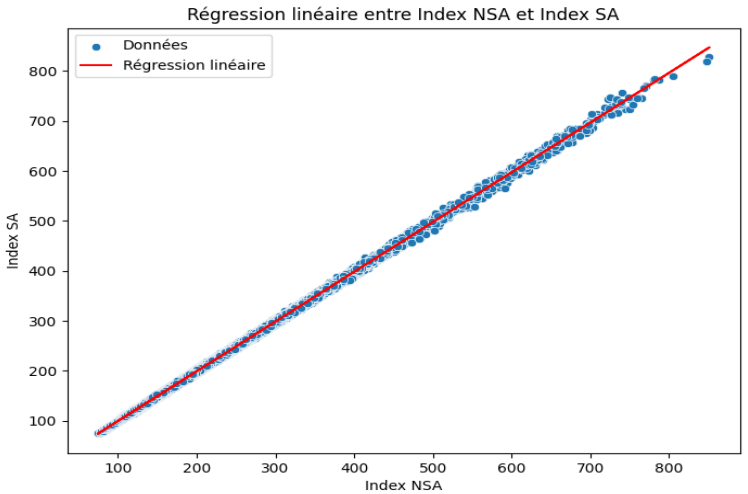

Figure 18 :Nuage de point et regression lineaire entre Index_nsa et index_sa

A.2 L’étude entre les années 'years' et 'index_nsa' :

La corrélation entre les années (yr) et l'indice NSA (index_nsa) montre une relation modérée mais positive, avec un coefficient de corrélation de 0,715. Cela suggère une tendance à l'augmentation de l'indice NSA avec le temps, bien que cette relation ne soit pas une dépendance linéaire forte.

La suppression des valeurs manquantes dans les colonnes 'index_nsa' et 'yr' a laissé un ensemble de données presque inchangé, passant de 123 970 à 123 969 enregistrements. Cependant, la régression linéaire entre les années et l'indice SA (seasonally adjusted) montre un p-value nul, indiquant une relation significative entre ces variables. Le coefficient de pente de la régression linéaire est de 1.72, ce qui suggère une augmentation de l'indice SA en fonction des années.

Le nuage de points associant les années à l'indice NSA confirme visuellement cette relation, montrant une tendance à l'augmentation de l'indice NSA au fil du temps, bien que la dispersion des données indique également une certaine variabilité autour de cette tendance**.**

Matrice de corrélation :

	yr	index_nsa
yr	1.00000	0.71542
index_nsa	0.71542	1.00000

Représentation graphique :

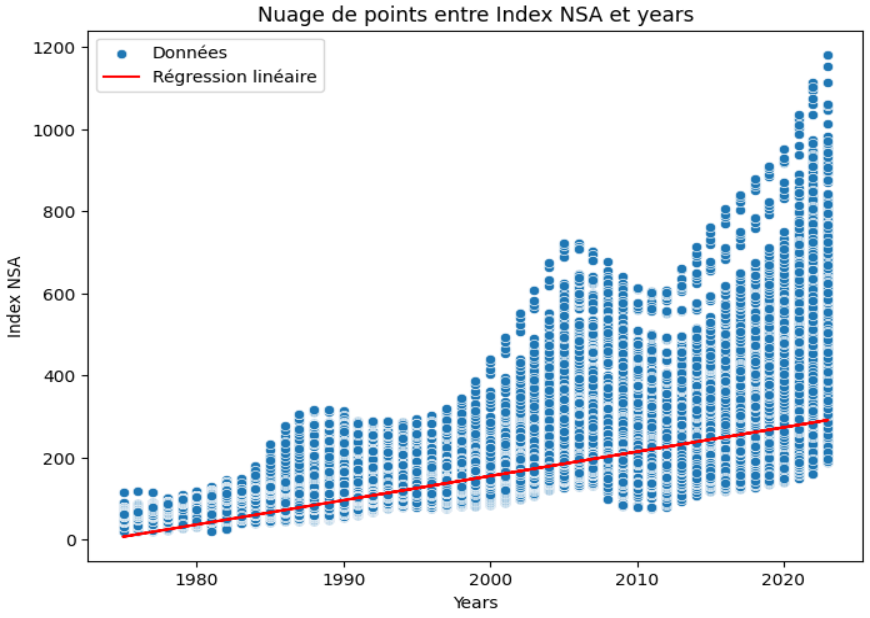

Figure 19 : Nuage de point et régression linéaire entre Index_nsa et les années

B.Variables Quantitatives-Qualitatives :

B.1 L’étude entre les années 'hpi_type'et 'index_nsa' :

Le graphique à boîtes moustache en montre la distribution de l'indice NSA pour chaque catégorie de 'hpi_type'. On constate des variations importantes entre les différentes catégories :

'distress-free' présente la médiane la plus élevée avec 197.87, illustrant une tendance générale à des valeurs plus élevées pour cette catégorie.

En revanche, 'traditional' affiche la médiane la plus basse, s'établissant à 158.52, démontrant une tendance vers des valeurs plus basses pour cette catégorie.

Le test de Student confirme des différences significatives entre les catégories de 'hpi_type' en ce qui concerne l'indice NSA.

les tests de Student permettent de comparer les moyennes de l'Index NSA entre les différentes catégories de HPI Type. Ces tests révèlent des différences significatives :

Entre 'traditional' et 'distress-free': T-statistic = -14.30, P-value = 9.10e-44

Entre 'traditional' et 'developmental': T-statistic = 2.61, P-value = 0.0096

Ces résultats confirme des différences significatives entre les catégories de 'hpi_type' en ce qui concerne l'indice NSA.

```
Moyenne de Index NSA pour traditional: 177.41069145656343
Moyenne de Index NSA pour non-metro: 183.87769800148038
Moyenne de Index NSA pour distress-free: 217.89447837150126
Moyenne de Index NSA pour developmental: 172.33506666666662


Test de Student entre traditional et non-metro: T-statistic = nan, P-value = nan
Test de Student entre traditional et distress-free: T-statistic = -14.30324890260876, P-value = 9.096013112069476e-44
Test de Student entre traditional et developmental: T-statistic = 2.6111574402902056, P-value = 0.009606393005279453
```

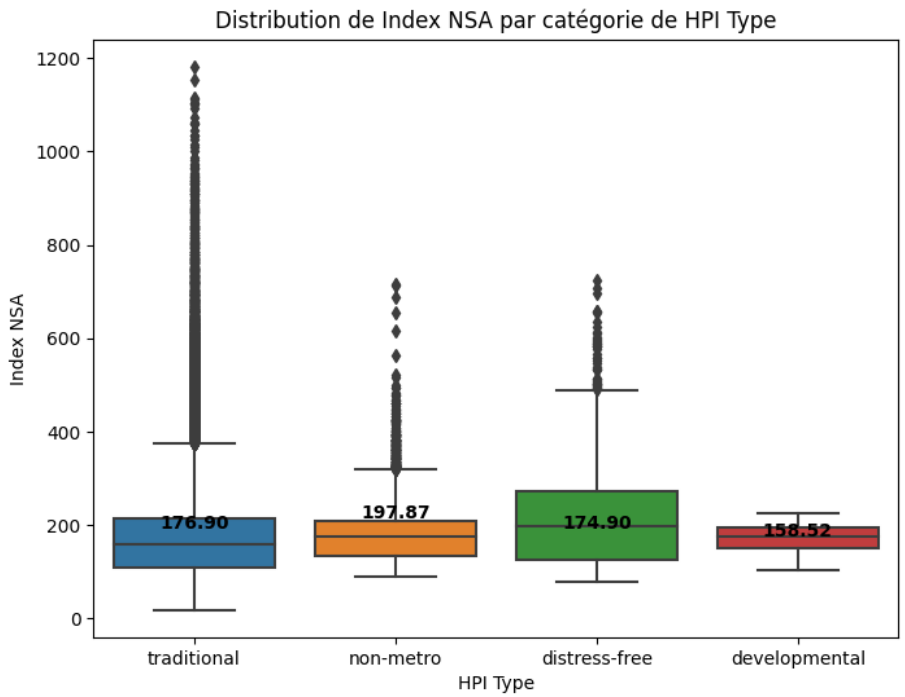

Figure 20:boîtes moustache en montre la distribution de l'indice NSA pour chaque catégorie de 'hpi_type'

B.2 L'étude entre "level"et "index_nsa":

Il exicte des différences significatives dans les distributions de l'indice NSA entre les différentes catégories de la variable 'level'.

Pour 'USA or Census Division', la médiane de l'indice NSA est de 149.13.

Pour 'MSA', la médiane est de 176.90.

Pour 'State', la médiane est de 188.66.

Pour 'Puerto Rico', la médiane est de 192.31.

Le test de Student effectué entre les catégories 'USA or Census Division' et 'MSA' montre une différence significative avec un t-statistic de 41.38 et un p-value très faible (proche de zéro). Cela suggère des différences statistiquement significatives entre ces deux catégories en termes de valeurs d'indice NSA.

Le test entre 'USA or Census Division' et 'Puerto Rico' montre également des différences significatives, avec un t-statistic de 17.76 et un p-value très faible.

Globalement, ces résultats suggèrent des variations substantielles des indices NSA entre les différents niveaux géographiques, en particulier entre 'USA or Census Division' et les autres catégories.

```
Médiane pour level: 149.13
Médiane pour level: 176.90
Médiane pour level: 188.66
Médiane pour level: 192.31


Moyenne de Index NSA pour USA or Census Division: 212.40386470588237
Moyenne de Index NSA pour MSA: 162.17121040527888
Moyenne de Index NSA pour State: 216.4154024589878
Moyenne de Index NSA pour Puerto Rico: 172.33506666666662


Test de Student entre USA or Census Division et MSA: T-statistic = 41.37605190565189, P-value = 0.0
Test de Student entre USA or Census Division et State: T-statistic = nan, P-value = nan
Test de Student entre USA or Census Division et Puerto Rico: T-statistic = 17.756758842046285, P-value = 3.9099992877045233e-53
```

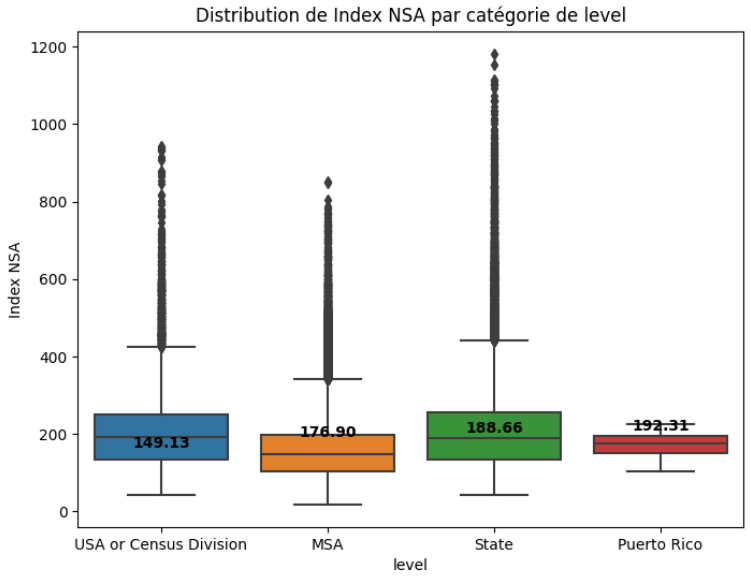

Figure 21:boîtes moustache en montre la distribution de l'indice NSA pour chaque catégorie de 'level'

B.3 L'étude entre "year"et "hpi_type":

la manière dont la variable yr varie selon différentes catégories de la variable hpi_type. Voici ce que les résultats signifient :

Le t-statistic négatif (-7,93) suggère que la valeur moyenne de yr pour la catégorie "non-metro" est significativement inférieure à la moyenne générale de yr. Cela peut indiquer qu'il y a un aspect temporel associé à la catégorie "non-metro" qui est notablement antérieur à la tendance globale.

Le t-statistic positif (12,12) suggère que la valeur moyenne de yr pour la catégorie "distress-free" est significativement supérieure à la moyenne générale de yr. Cela peut indiquer qu'il y a un aspect temporel associé à la catégorie "distress-free" qui est notablement postérieur à la tendance globale.

Similaire à la catégorie "distress-free", le t-statistic positif (9,58) suggère que la valeur moyenne de yr pour la catégorie "developmental" est significativement supérieure à la moyenne générale de yr. Cela peut indiquer un aspect temporel associé à la catégorie "developmental" qui est notablement postérieur à la tendance globale.

yr for hpi_type - Category: traditional

T-Statistic: -7.925992163463643, Mean: 2003.710203137846

yr for hpi_type - Category: non-metro

T-Statistic: 43.33344304655495, Mean: 2008.878260869565

yr for hpi_type - Category: distress-free

T-Statistic: 12.124870702285541, Mean: 2006.8778625954199

yr for hpi_type - Category: developmental

T-Statistic: 9.578403729828489, Mean: 2009.1866666666667

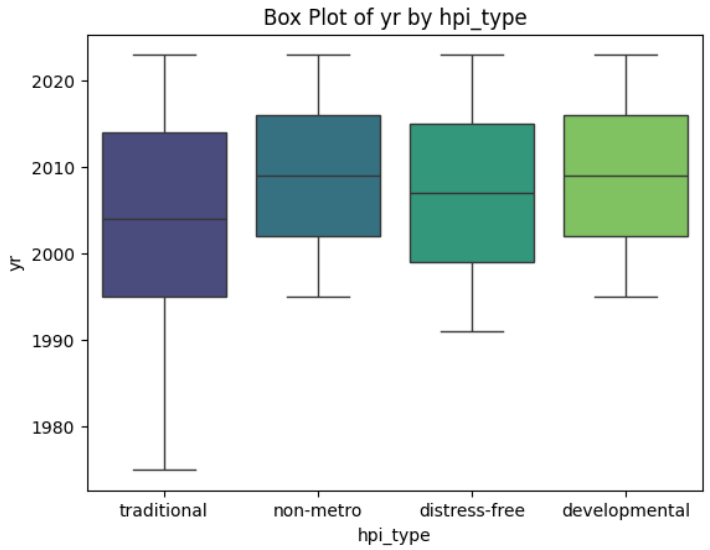

Figure 22:boîtes moustache en montre la distribution de l'indice year pour chaque catégorie de 'hpi_type'

B.4 L'étude entre "year"et "hpi_flavor":

Le t-statistic positif (50,19) suggère que la valeur moyenne de yr pour la catégorie "purchase-only" est significativement supérieure à la moyenne générale de yr. Cela peut indiquer qu'il y a un aspect temporel associé à la catégorie "purchase-only" qui est notablement postérieur à la tendance globale.

Le t-statistic négatif (-33,23) suggère que la valeur moyenne de yr pour la catégorie "all-transactions" est significativement inférieure à la moyenne générale de yr. Cela peut indiquer qu'il y a un aspect temporel associé à la catégorie "all-transactions" qui est notablement antérieur à la tendance globale.

Le t-statistic positif (36,89) suggère que la valeur moyenne de yr pour la catégorie "expanded-data" est significativement supérieure à la moyenne générale de yr. Cela peut indiquer qu'il y a un aspect temporel associé à la catégorie "expanded-data" qui est notablement postérieur à la tendance globale.

yr for hpi_flavor - Category: purchase-only

T-Statistic: 50.19017831261321, Mean: 2006.8882564409826

yr for hpi_flavor - Category: all-transactions

T-Statistic: -33.23101392236212, Mean: 2002.5402719854942

yr for hpi_flavor - Category: expanded-data

T-Statistic: 36.886820877296636, Mean: 2006.8778625954199

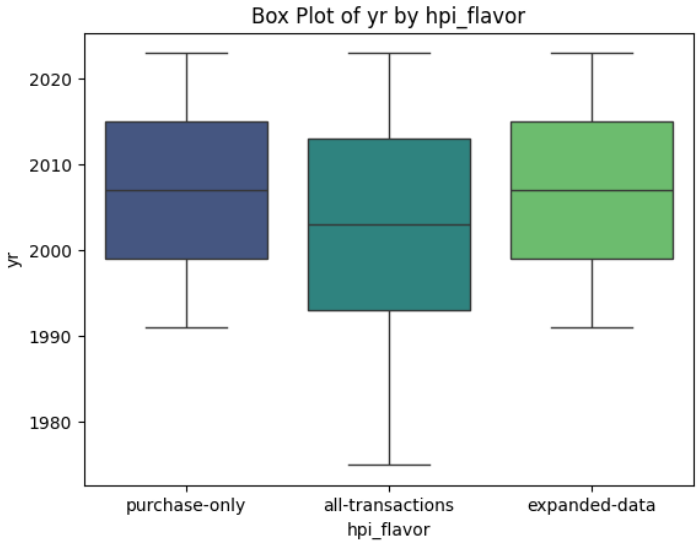

Figure 23:boîtes moustache en montre la distribution de l'indice year pour chaque catégorie de 'hpi_flavor'

B.5 L'étude entre "year"et "level":

L'année moyenne (yr) pour la catégorie "USA ou Division du recensement" est significativement plus tardive que la moyenne globale (Moyenne: 2005.04). Cela suggère que, en moyenne, les tendances ou événements au sein des États-Unis ou des divisions du recensement ont tendance à se produire plus tard que la tendance générale.

L'année moyenne (yr) pour la catégorie "MSA" est significativement antérieure à la moyenne globale (Moyenne: 2003.70). Cela indique que, en moyenne, les tendances ou événements au sein des zones statistiques métropolitaines ont tendance à se produire plus tôt que la tendance globale.

L'année moyenne (yr) pour la catégorie "État" est significativement plus tardive que la moyenne globale (Moyenne: 2004.48). Cela suggère que, en moyenne, les tendances ou événements au sein des États individuels ont tendance à se produire plus tard que la tendance globale.

L'année moyenne (yr) pour la catégorie "Puerto Rico" est significativement plus tardive que la moyenne globale (Moyenne: 2009.19). Cela implique que, en moyenne, les tendances ou événements à Porto Rico ont tendance à se produire plus tard que la tendance globale.

yr for level - Category: USA or Census Division

T-Statistic: 8.691340518786122, Mean: 2005.0423529411764

yr for level - Category: MSA

T-Statistic: -7.051983772355398, Mean: 2003.7032346041394

yr for level - Category: State

T-Statistic: 7.115442128896252, Mean: 2004.4831429367512

yr for level - Category: Puerto Rico

T-Statistic: 9.578403729828489, Mean: 2009.1866666666667

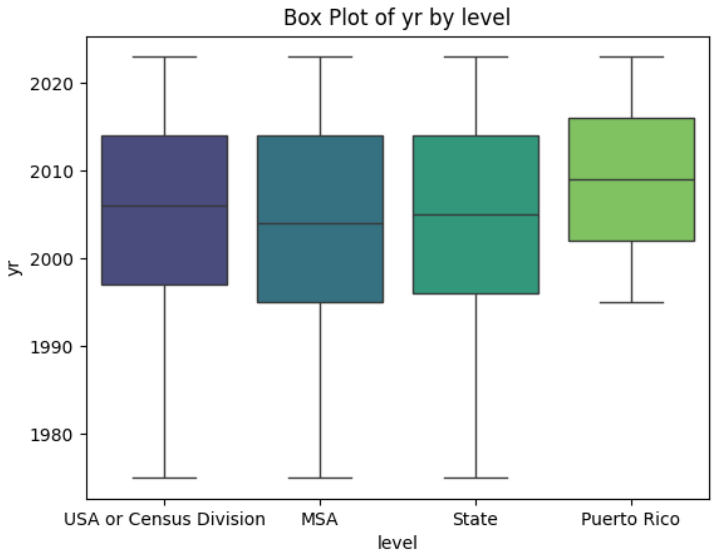

Figure 24:boîtes moustache en montre la distribution de l'indice year pour chaque catégorie de 'level'

C.Variables Qualitatives:

C.1 L'étude entre " hpi_type "et "hpi_flavor":

L'analyse de la variable 'hpi_type' et la variable 'hpi_flavor' montre une association significative entre ces deux variables.

La table de contingence révèle la répartition des occurrences pour chaque paire de modalités :

Il y a 114 occurrences pour 'developmental - all-transactions' et 111 pour 'developmental - purchase-only'. Aucune occurrence n'est présente pour 'distress-free - expanded-data'. En revanche, 'distress-free - purchase-only' a 1572 occurrences. 'Non-metro - all-transactions' totalise 5405 occurrences, mais aucune pour 'non-metro - expanded-data' ou 'non-metro - purchase-only'. 'Traditional - all-transactions' est représenté par 77206 occurrences, suivi de 'traditional - expanded-data' avec 14541 occurrences et 'traditional - purchase-only' avec 25021 occurrences.

Le test du Chi-deux (Chi-square) indique une statistique de Chi2 de 8595.73 avec un p-value très proche de zéro (0.0). Cela suggère une association significative entre les variables 'hpi_type' et 'hpi_flavor', ce qui signifie que ces variables ne sont probablement pas indépendantes et qu'il existe une relation entre les différentes modalités de ces deux variables.

```
Chi2 Statistic: 8595.729984916265
P-value: 0.0
```

hpi_flavor hpi_type	all-transactions	expanded-data	purchase-only
developmental	114	0	111
distress-free	0	0	1572
non-metro	5405	0	0
traditional	77206	14541	25021

Figure 25: Tableau de contingence

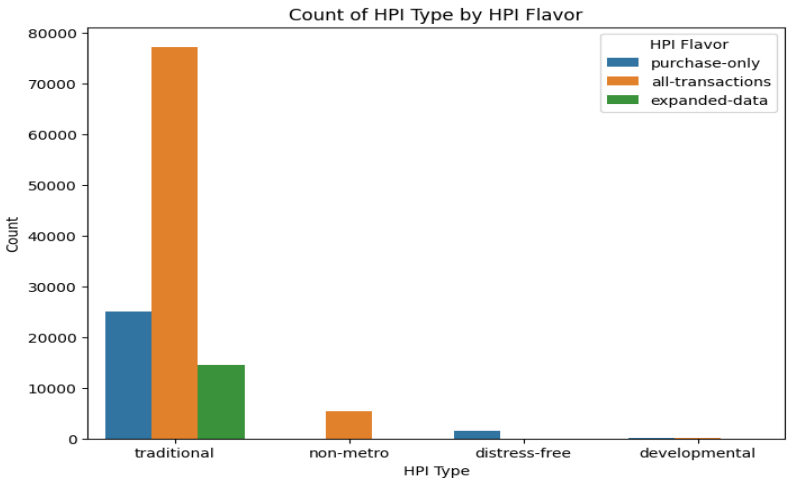

Figure 26: Nombre de 'hpi_type' par 'hpi_flavor'

4. Analyse Multivariée :

L'analyse multivariée s'est concentrée sur quatre variables spécifiques ('yr', 'period', 'index_nsa', 'index_sa') parmi les dix disponibles dans les données. Ce choix a été motivé par la pertinence de ces variables pour étudier les tendances temporelles des prix des maisons. En sélectionnant ces variables clés, l'objectif était de comprendre les corrélations potentielles entre les informations temporelles et les indices de prix des maisons, simplifiant ainsi l'analyse tout en se concentrant sur des indicateurs cruciaux pour l'étude des fluctuations des prix immobiliers.

Dans cette section d'analyse multivariée, la corrélation entre les variables est significative :

- Il existe une forte corrélation entre 'index_nsa' et 'index_sa' avec des valeurs de corrélation de 0.999739.

- Les variables 'yr' et 'index_sa' présentent également une corrélation notable de 0.768658, tandis que 'yr' et 'index_nsa' ont une corrélation de 0.715420.

- Le calcul des moyennes des variables ('yr', 'period', 'index_nsa', 'index_sa') montre des valeurs respectives de 2003.99, 2.62, 178.20, et 199.67.

De plus, l'analyse en composantes principales (PCA) a été utilisée pour réduire la dimensionnalité des données. Cette analyse a permis d'obtenir deux composantes principales expliquant respectivement environ 52% et 25% de la variance totale. Les composantes principales ont été identifiées, mettant en lumière les contributions des différentes variables à ces composantes.

Pour visualiser l'impact de chaque composante, une représentation graphique a été réalisée en utilisant la méthode du coude (Elbow Method), montrant la décroissance

des valeurs propres en fonction du nombre de composantes principales. Cette méthode permet d'identifier le nombre optimal de composantes à retenir pour expliquer au mieux la variance des données.

	PC1	PC2	PC3	PC4
yr	8.298350	-1.821066	-8.134094	0.004042
period	0.025341	-0.024140	0.023458	1.435722
index_nsa	96.331058	-14.434417	0.739161	-0.000984
index_sa	35.456043	39.643316	-0.104502	0.000702

Figure27:La corrélation entre les variables initiales et les axes

	yr	period	index_nsa	index_sa
yr	1.000000	0.004099	0.715420	0.768658
period	0.004099	1.000000	0.020052	-0.001028
index_nsa	0.715420	0.020052	1.000000	0.999739
index_sa	0.768658	-0.001028	0.999739	1.000000

Figure28: Matrice corrélation

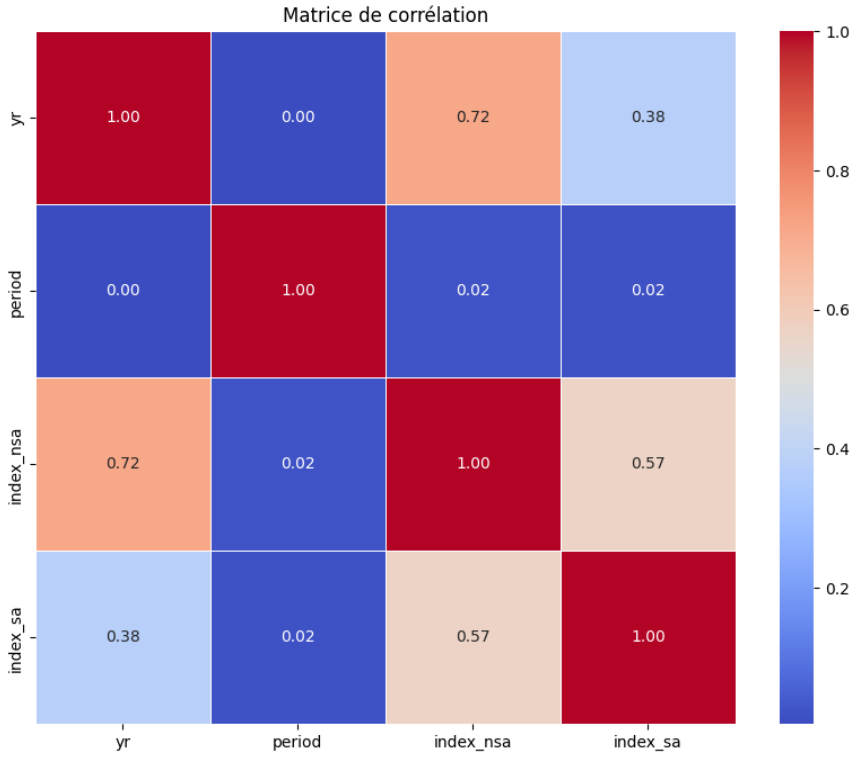

Figure29: Matrice corrélation

	yr	period	index_nsa	index_sa
yr	138.342387	0.069243	819.669029	669.931448
period	0.069243	2.063072	2.805549	-0.180174
index_nsa	819.669029	2.805549	9488.647942	8545.976218
index_sa	669.931448	-0.180174	8545.976218	8502.457926

Figure30: Matrice covariance

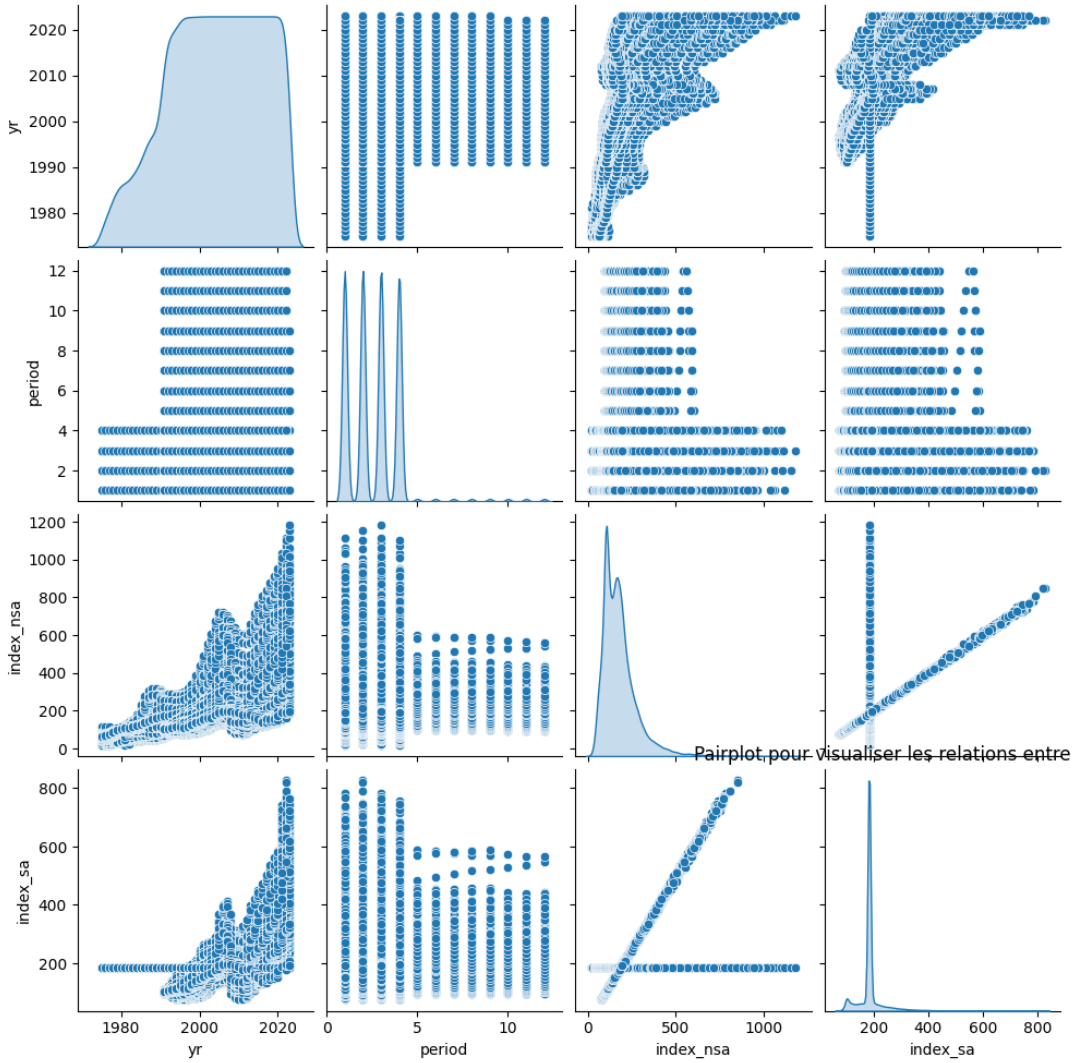

Figure31: Pairplot pour visualiser les relations entre les variables

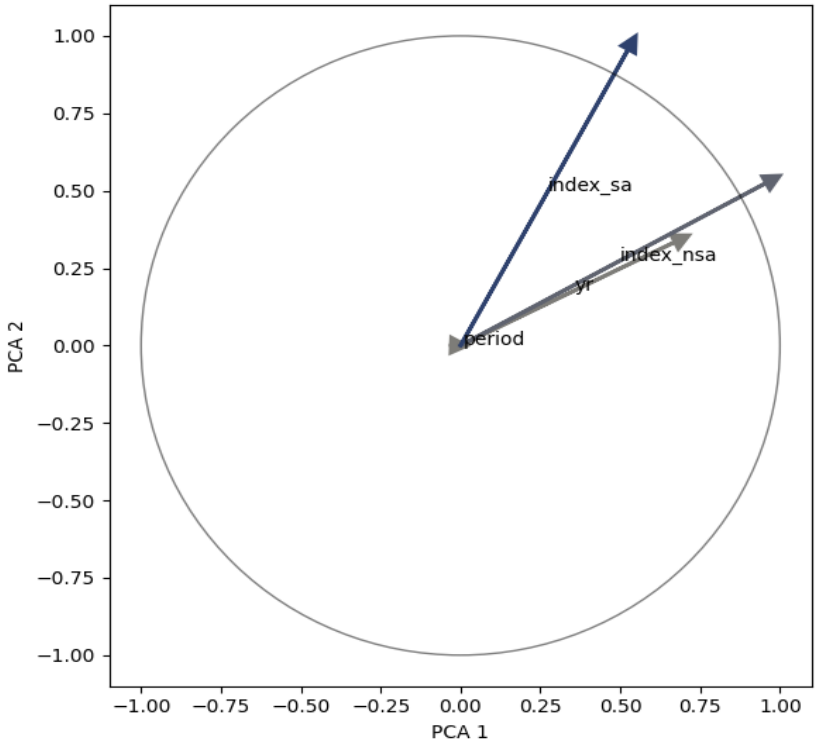

Figure 32:cercle de corrélation

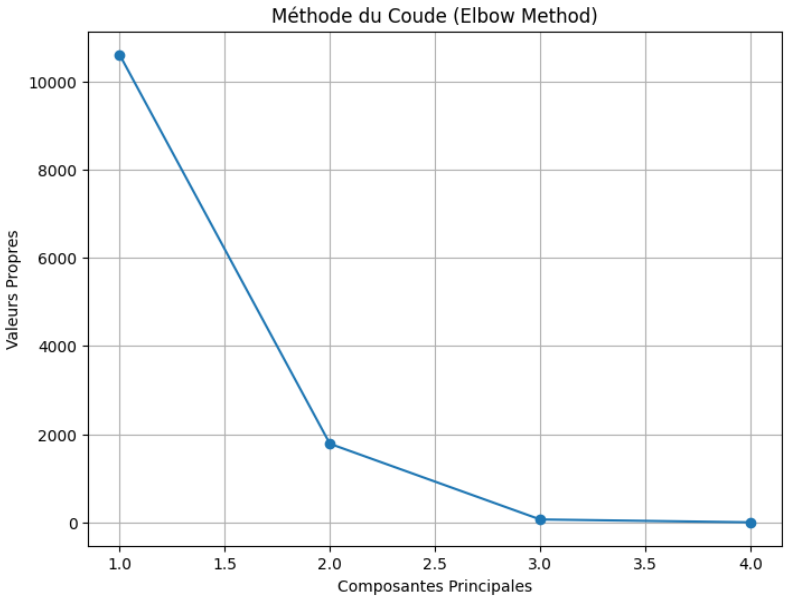

Figure 33:Methode du coude(Elbow Method)

5. conclusion:

Dans cette étude, l'analyse univariée a exploré différentes variables telles que 'hpi_type', 'hpi_flavor', 'level', 'place_name', 'yr', 'index_nsa' et 'index_sa'. Chacune de ces variables a été examinée individuellement pour comprendre leurs distributions, leurs fréquences et leurs corrélations potentielles avec d'autres facteurs.

Dans l'analyse bivariée, des corrélations intéressantes ont émergé. Par exemple, il a été observé une forte corrélation entre 'index_nsa' et 'index_sa', ainsi qu'une corrélation modérée mais positive entre 'yr' et 'index_nsa'. Ces relations entre les variables ont été étudiées plus en détail pour en extraire des tendances significatives.

Enfin, l'analyse multivariée a concentré l'attention sur quatre variables spécifiques - 'yr', 'period', 'index_nsa', 'index_sa' - révélant des corrélations significatives entre elles. L'utilisation de méthodes comme l'analyse en composantes principales (PCA) a permis de réduire la dimensionnalité des données et d'identifier les composantes clés expliquant la variance dans l'ensemble des variables étudiées.

Ces analyses ont fourni une compréhension approfondie des relations entre les variables étudiées, mettant en lumière des tendances, des corrélations et des distributions pertinentes pour mieux appréhender les fluctuations des prix des logements et leurs relations avec d'autres facteurs temporels et géographiques.