ANALYSE DE DONNÉES

ANALYSE DE L'INDICE DES PRIX DES MAISONS (HPI)

SUPERVISÉ PAR :

Dr. M.FAREH

ANALYSE DE L'INDICE DES PRIX DES MAISONS (HPI)

HOUSE PRICE INDEXES

REALISÉ ET PRESENTÉ PAR :

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

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. Avec 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.

INTRODUCTION

En explorant les trois types d'analyse, chaque étape s'inscrit dans une méthodologie rigoureuse visant à analysé les secrets et les tendances au sein de notre ensemble de données.

OBJECTIFS ET RAISONS :

OBJECTIFS

- Compréhension des Tendances des Prix des Maisons.
- Analyse Géographique complète en examinant les fluctuations des prix des maisons à différents niveaux.
- Compréhension de la Méthodologie du HPI .
- Évaluer le HPI en tant qu'indicateur économique.
- Fournir une description détaillée de l'ensemble de données choisi.

RAISONS

- Couverture Complète : Données disponibles pour les 50 États et plus de 400 villes aux États-Unis
- Données Longitudinales : Disponibles depuis les années 1970
- Méthodologie transparente basée sur des techniques statistiques fiables
- Fournit des insights précieux pour la Prise de Décision
- Outil analytique crucial pour estimer les tendances économiques régionales

EXAMPLE DE LA TABLE HPI_MASTER.CSV

hpi_type	hpi_flavor	frequency	level	place_name	place_id	yr	period	index_nsa	index_sa
traditional	purchase-only	monthly	USA or Census Division	East North Central Division	DV_ENC	**1991**	**1**	**100.00**	**100.00**
traditional	purchase-only	monthly	USA or Census Division	East North Central Division	DV_ENC	**1991**	**2**	**100.91**	**100.97**
traditional	purchase-only	monthly	USA or Census Division	East North Central Division	DV_ENC	**2023**	**9**	**342.02**	**337.13**
traditional	purchase-only	monthly	USA or Census Division	East South Central Division	DV_ESC	**2004**	**10**	**171.54**	**170.61**
….	….	…	…	….	…	…	…	…	…
traditional	all-transactions	quarterly	MSA	Columbus	OH	**2005**	**1**	**174.65**	**176.02**
….	….	…	…	….	…	…	…	…	…

ANALYSE UNIVARIÉE

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

LA VARIABLE 'HPI_TYPE'

'traditional' représentant environ 94% de l'ensemble des données.
'non-metro' environ 4.36%
'distress-free' environ 1.27%
'developmental' environ 0.18%

	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

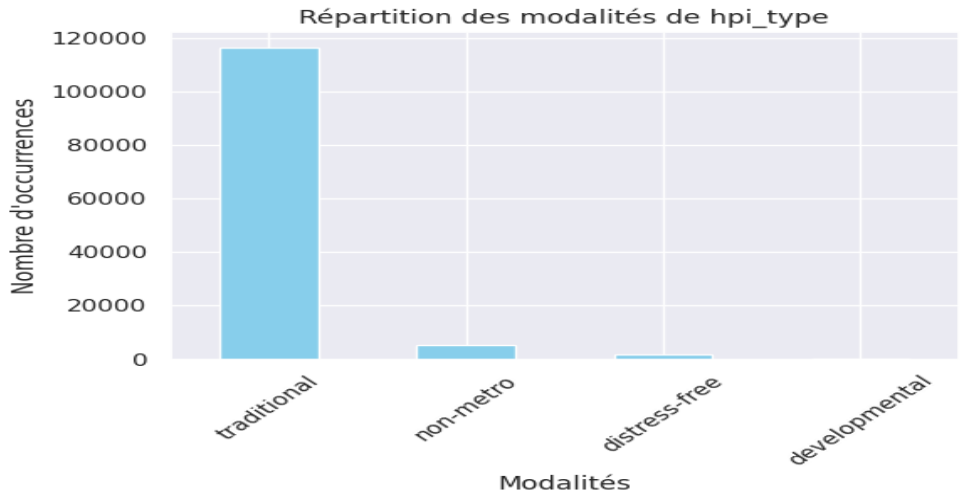
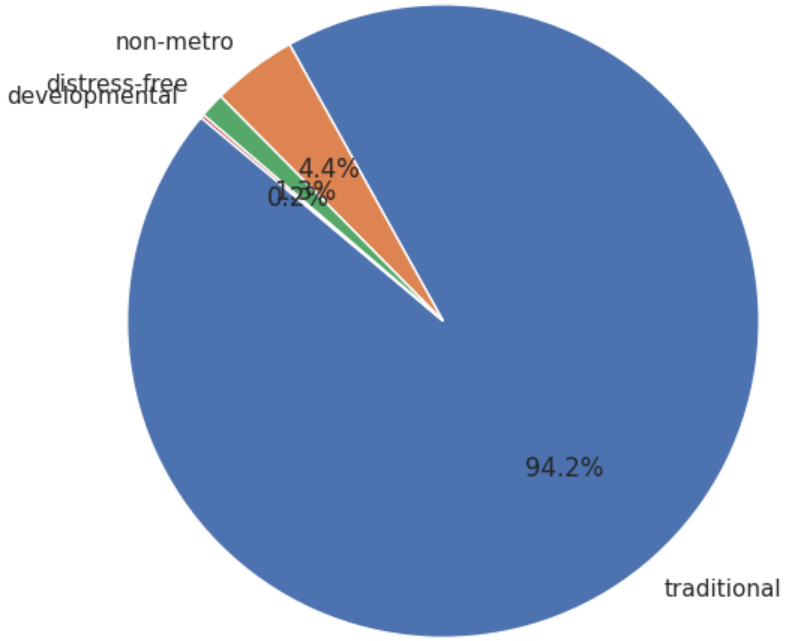


Diagramme de tuyaux de 'hpi_type'




```
traditional      94.190530
non-metro         4.359926
distress-free     1.268049
developmental     0.181496
```

Diagramme circulaire de 'hpi_type'

LA VARIABLE 'HPI_FLAVOR'

'all-transactions', présente environ 66.73% de l'ensemble des données.
'purchase-only' soit environ 21.54%.
'expanded-data' représentant environ 11.73%.

	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

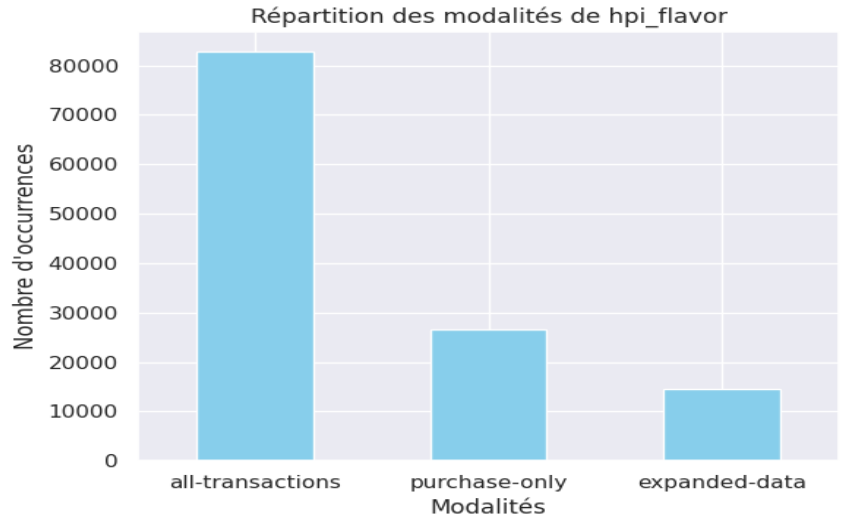


Diagramme de tuyaux de 'hpi_flavor'

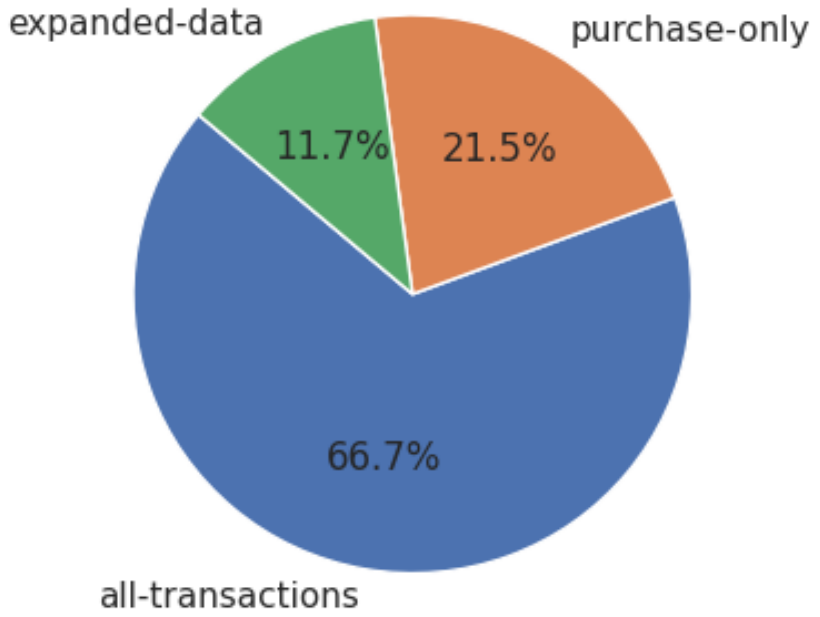
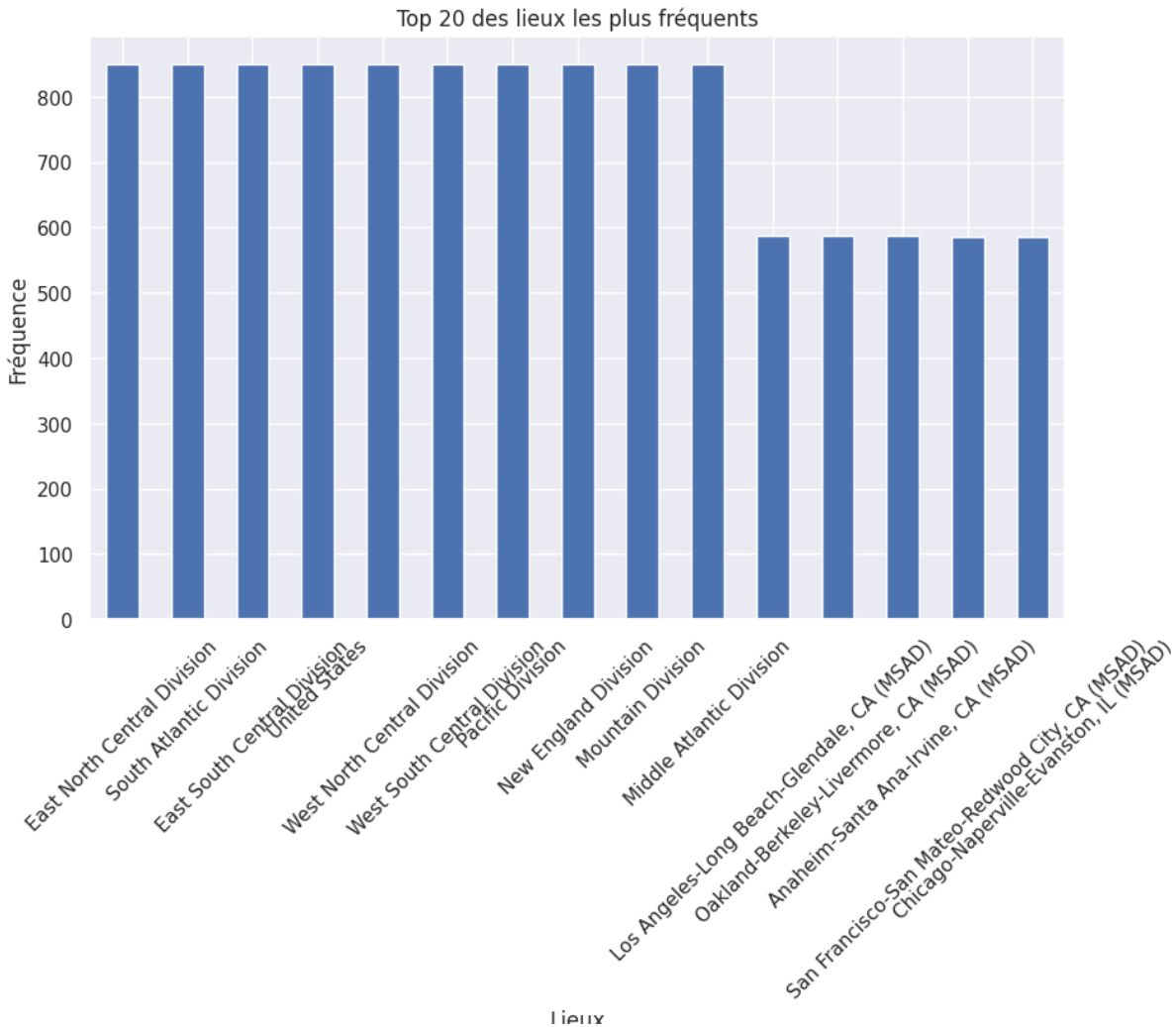
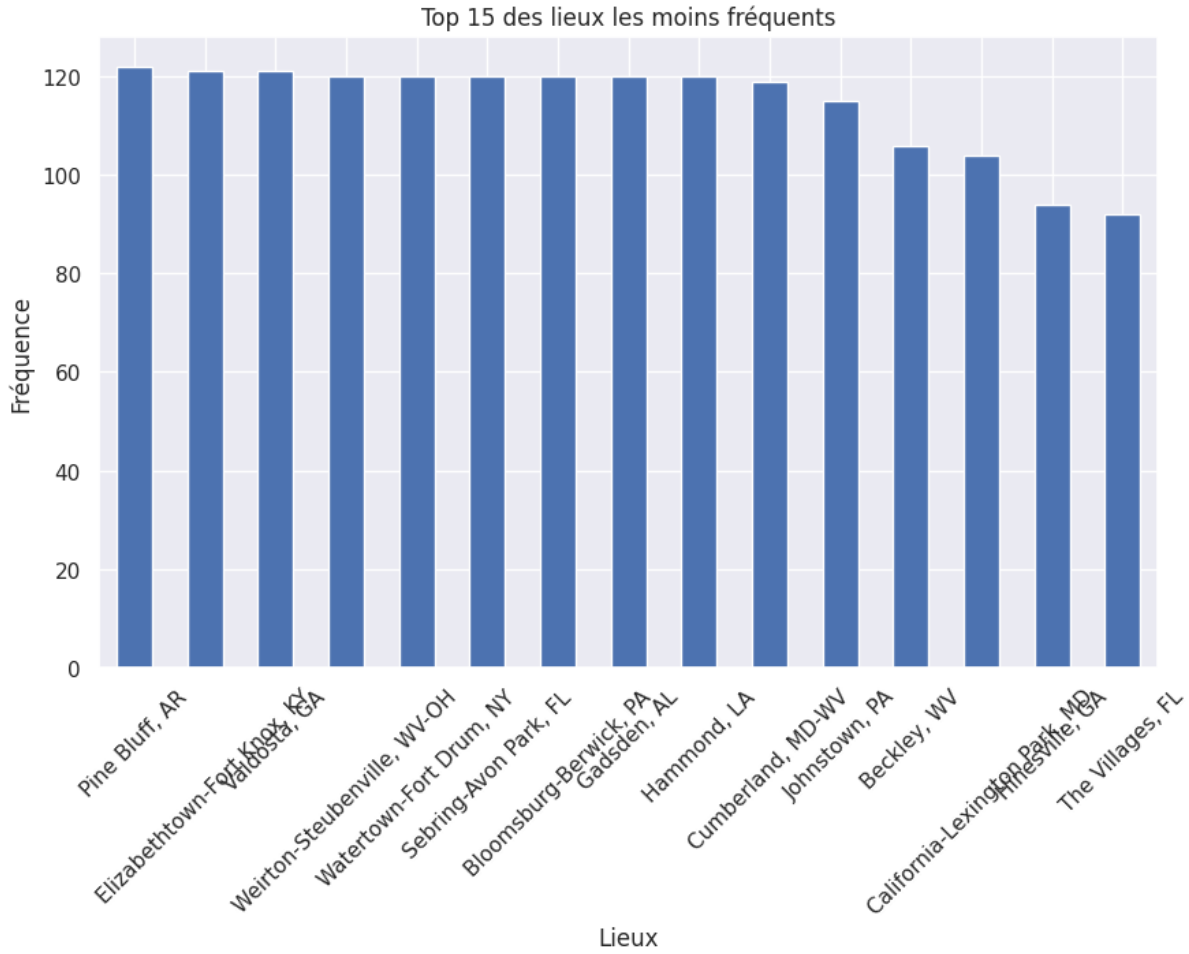


Diagramme circulaire de 'hpi_flavor'

LA VARIABLE 'PLACE_NAME'

- comporte un total de 466 valeurs uniques
- 10 premiers lieux ont tous une fréquence de 850, ce qui suggère une uniformité apparente dans ces catégories.





LA VARIABLE 'LEVEL'

'MSA' constitue la majorité avec 69.80%
'State' représente 23.16%
'USA or Census Division' à 6.86%
'Puerto Rico' à seulement 0.18%

	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

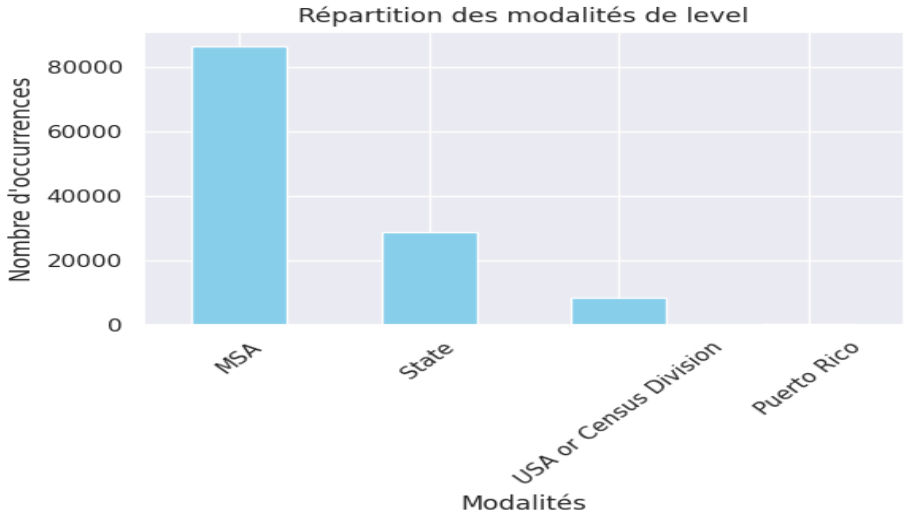


Diagramme de tuyaux de 'level'

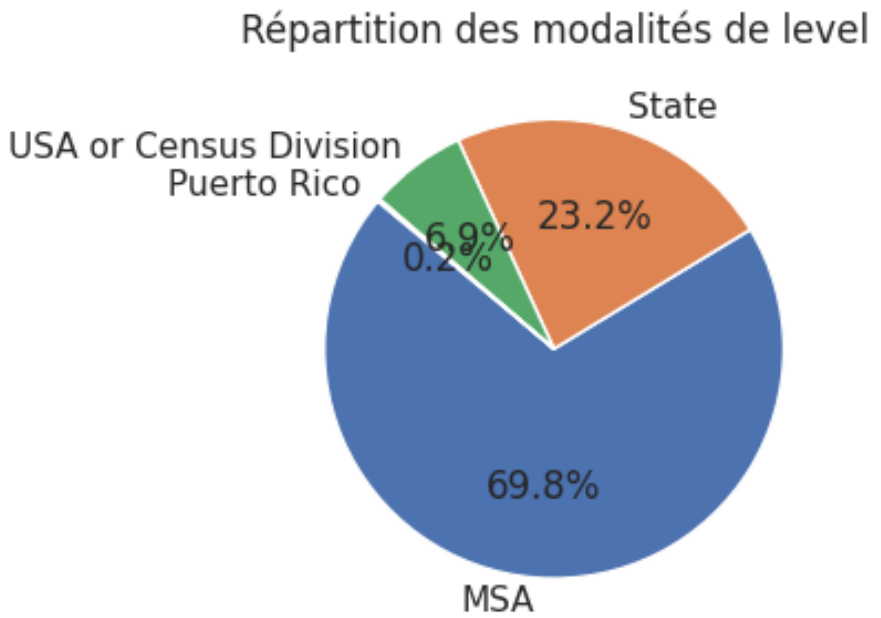


Diagramme circulaire de 'level'

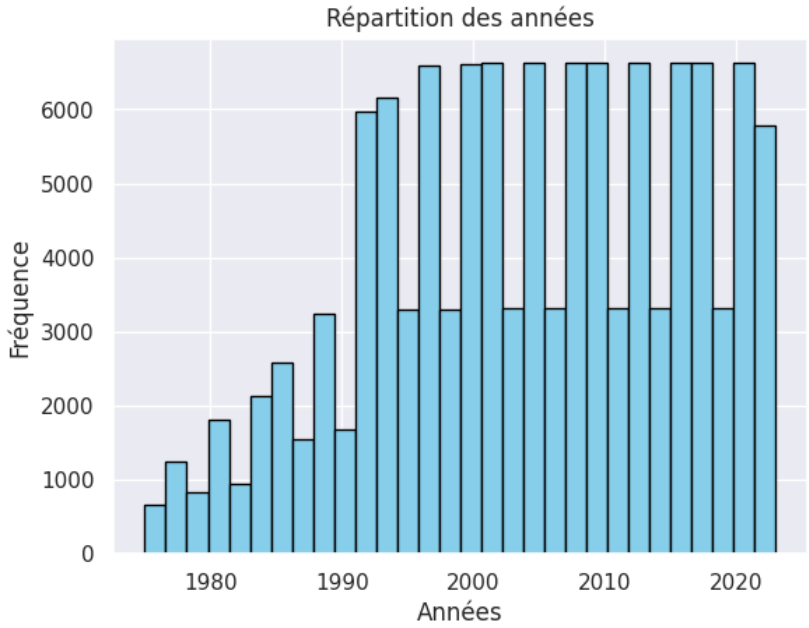
LA VARIABLE 'YEARS'

50% des données se situent entre 1995 et 2014
La médiane se trouve autour de 2005

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

Diagramme de tuyaux de 'years'



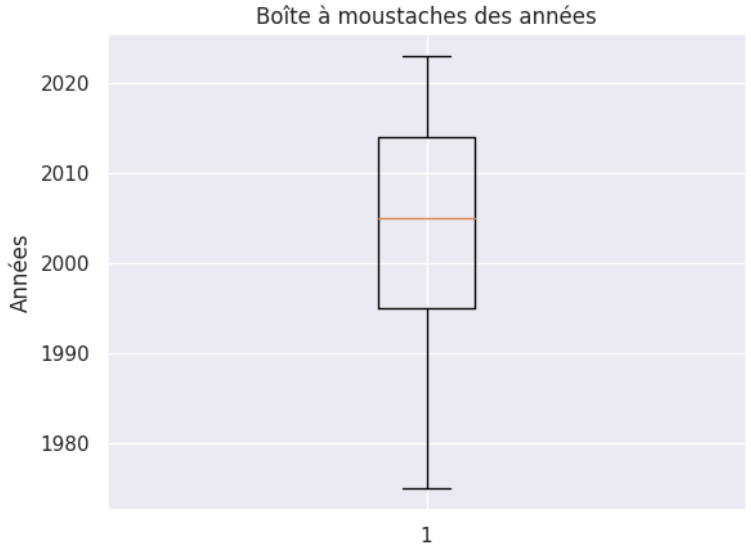


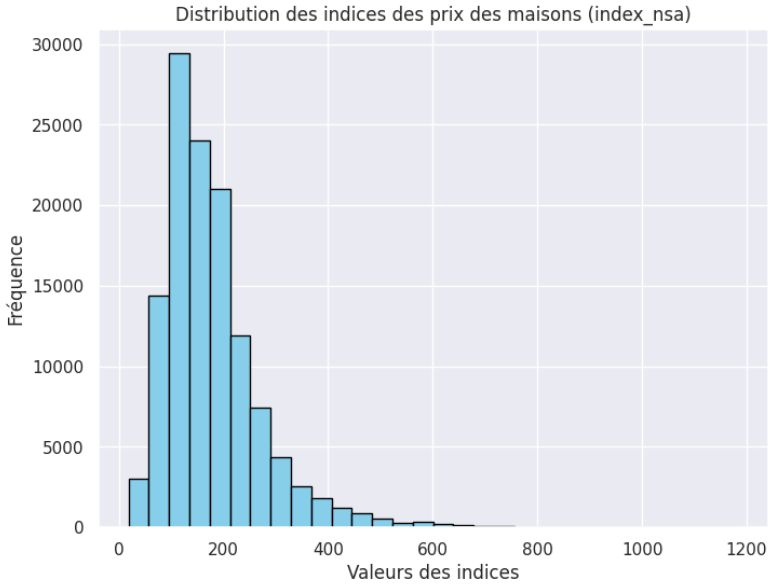
Diagramme a moustaches de 'years'

LA VARIABLE 'INDEX_NSA' (NOT SEASONALLY ADJUSTED)'

```
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
```

Diagramme de tuyaux de 'Index_NSA '



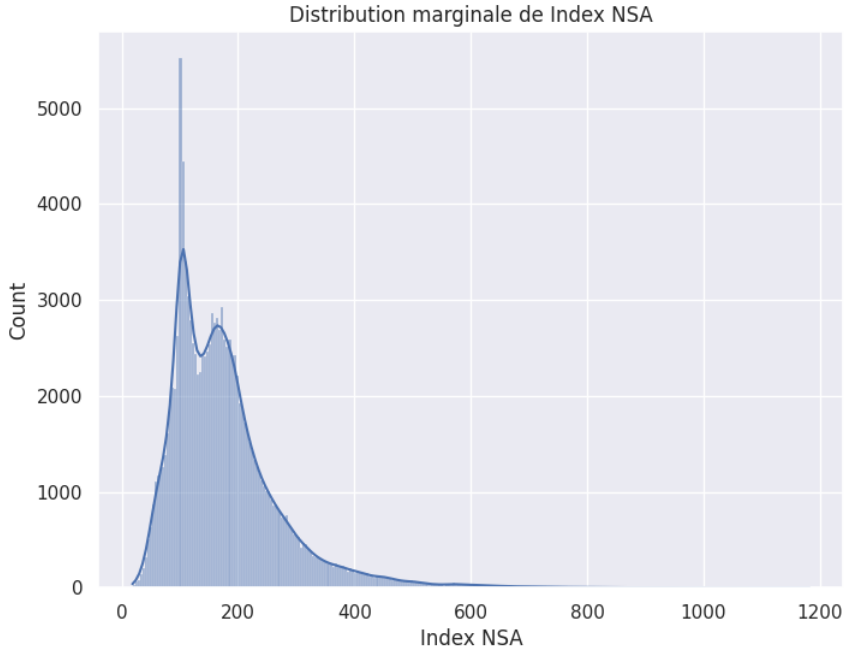


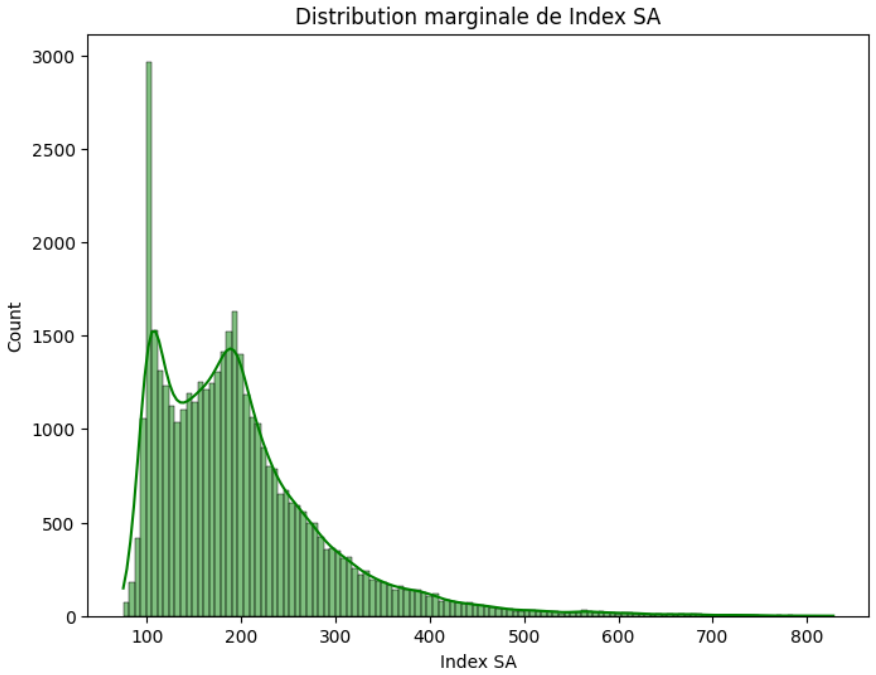
Diagramme a moustaches de 'Index_NSA '

LA VARIABLE 'INDEX_SA' (SEASONALLY ADJUSTED)

```
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

Diagramme de tuyaux de 'Index_SA '



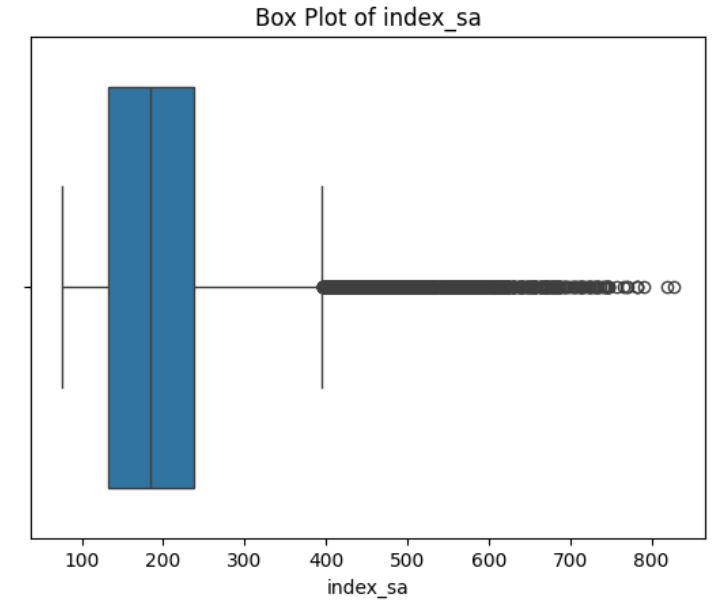


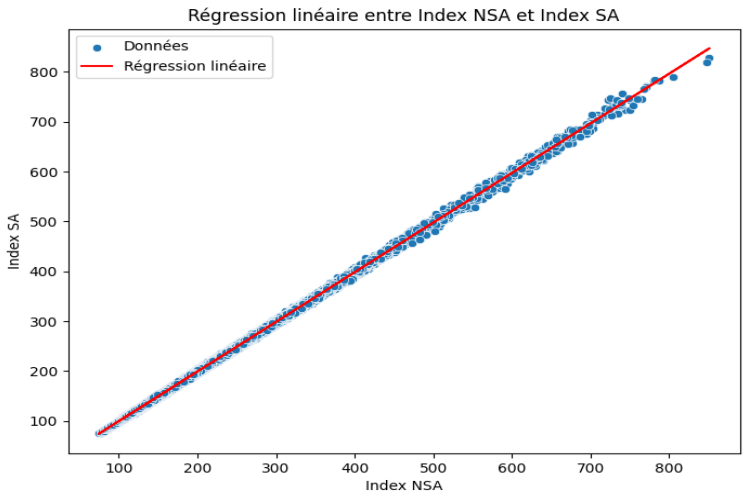
Diagramme a moustaches de 'Index_SA '

ANALYSE BIVARIÉE

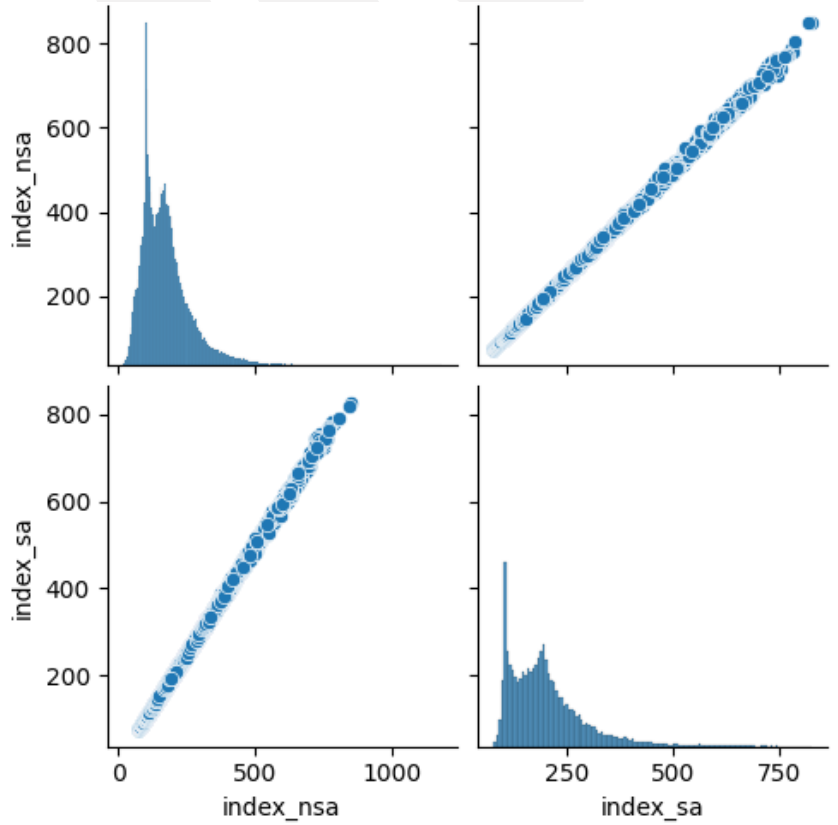
L’ÉTUDE ENTRE 'INDEX_SA ' ET 'INDEX_NSA'

Il existe une corrélation extrêmement forte entre les variables 'index_nsa' et 'index_sa'. La corrélation de 0,9997 indique une dépendance quasi totale .

	index_nsa	index_sa
index_nsa	1.000000	0.999739
index_sa	0.999739	1.000000



Nuage de point et regression lineaire entre Index_nsa et index_sa

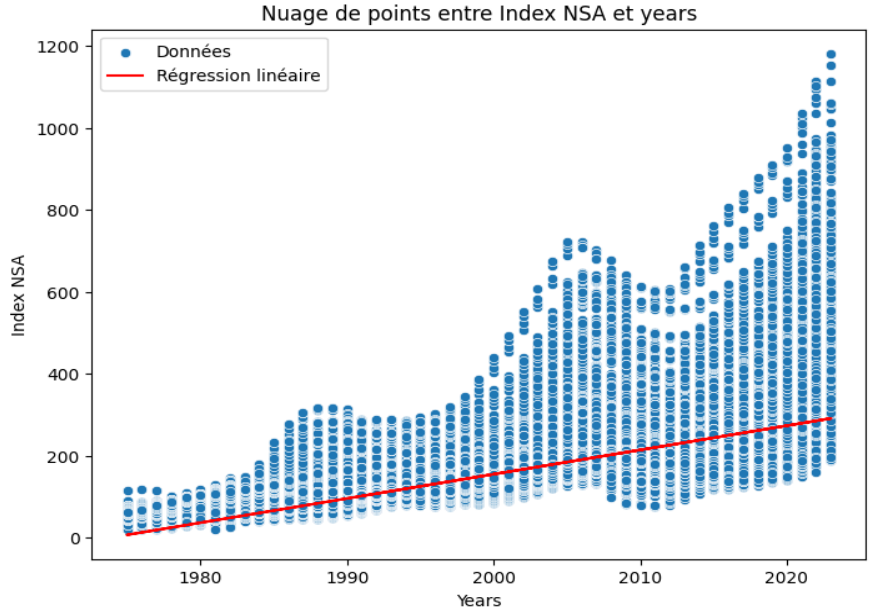


Nuage de point et Diagramme de tuyaux entre Index_nsa et index_sa

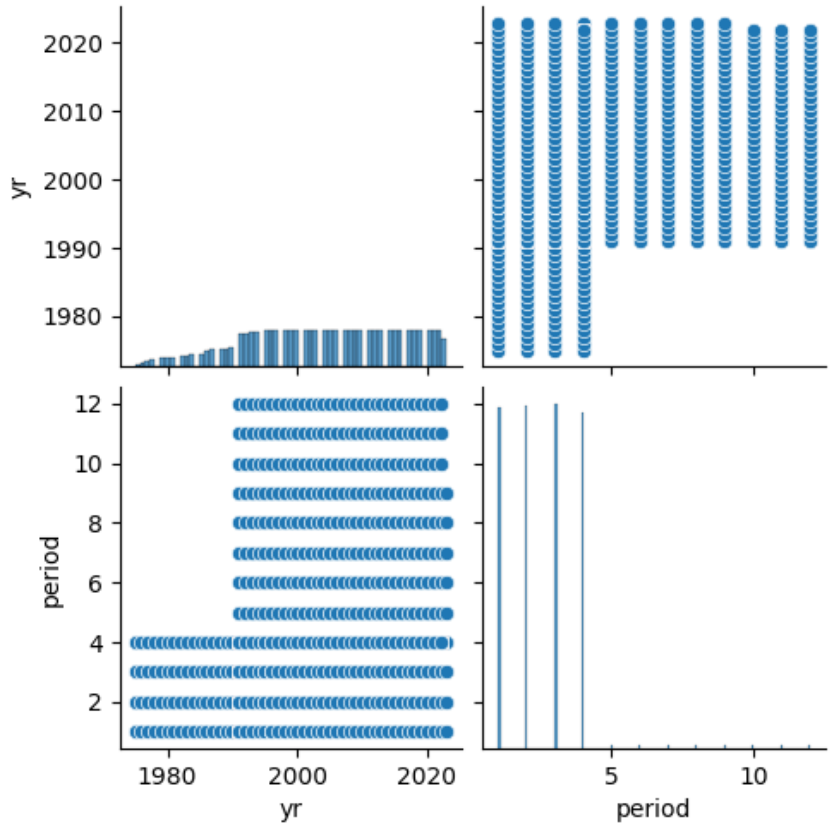
L'ÉTUDE ENTRE 'YEARS' ET 'INDEX_NSA'

La corrélation entre les années (yr) et l'indice NSA (index_nsa) montre une relation modérée mais positive.

	yr	index_nsa
yr	1.00000	0.71542
index_nsa	0.71542	1.00000



Nuage de point et regression lineaire entre Index_nsa et et 'years' (*les années)*

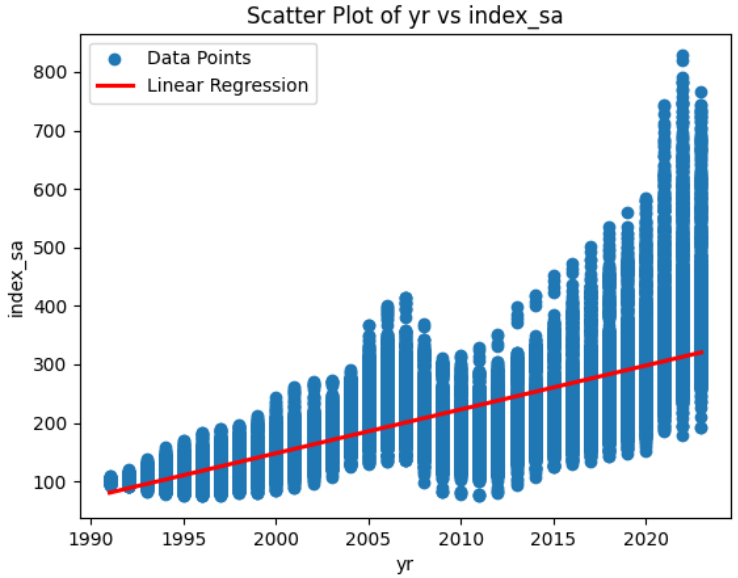


Nuage de point et Diagramme de tuyaux entre Index_nsa et 'years' (*les années)*

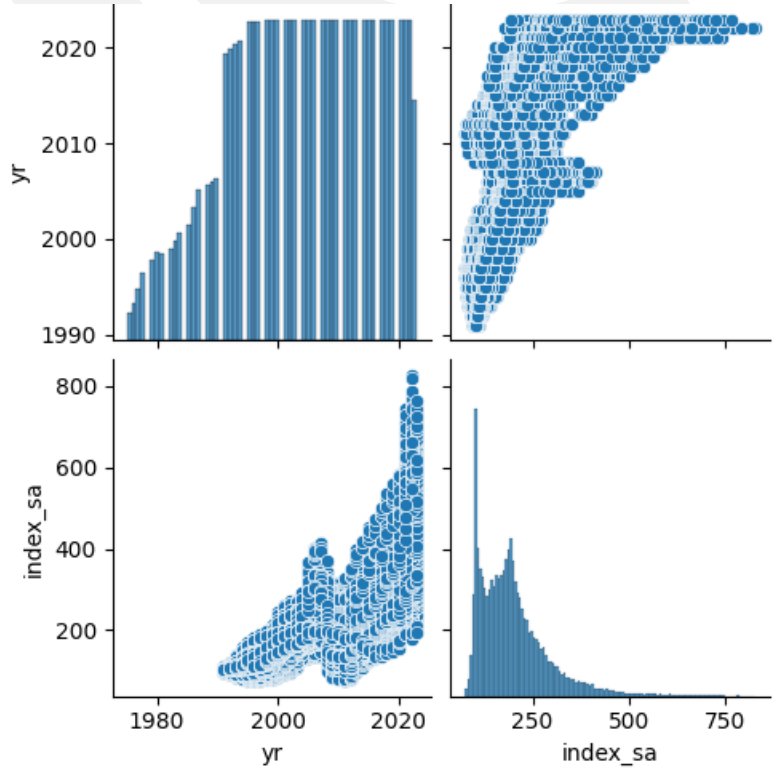
L'ÉTUDE ENTRE 'YEARS' ET 'INDEX_SA'

La corrélation entre les années (yr) et l'indice NSA (index_nsa) montre une relation modérée mais positive.

	yr	index_nsa
yr	1.00000	0.71542
index_nsa	0.71542	1.00000



Nuage de point et regression lineaire entre Index_sa et et 'years' (*les années)*

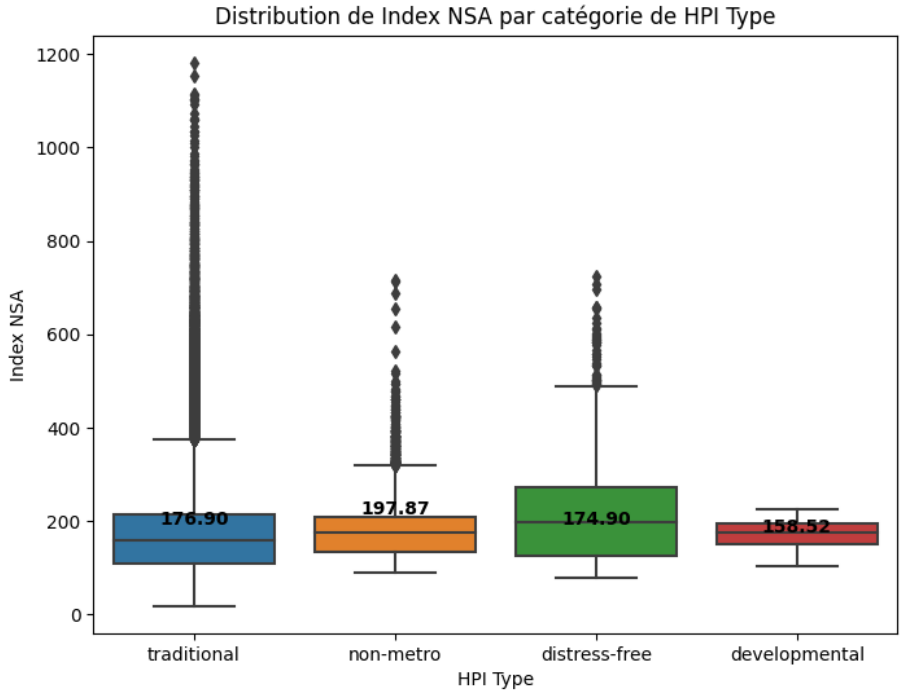


Nuage de point et Diagramme de tuyaux entre Index_sa et et 'years' (*les années)*

L’ÉTUDE ENTRE ‘HPI_TYPE 'ET 'INDEX_NSA' :

'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.

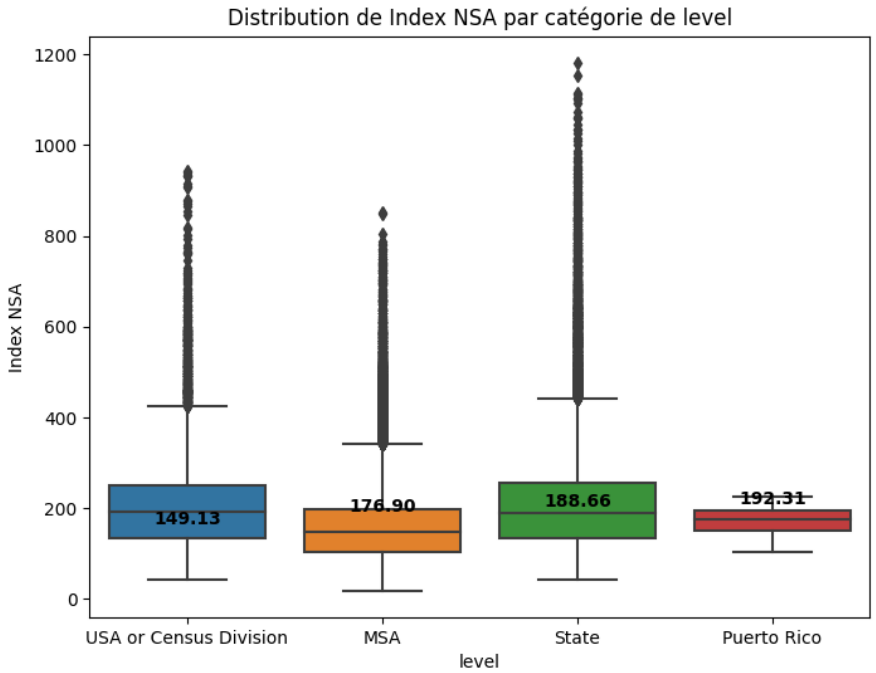


boîtes moustache en montre la distribution de l'indice NSA pour chaque catégorie de 'hpi_type'

L’ÉTUDE ENTRE ‘LEVEL’ ET ‘INDEX_NSA’:

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.



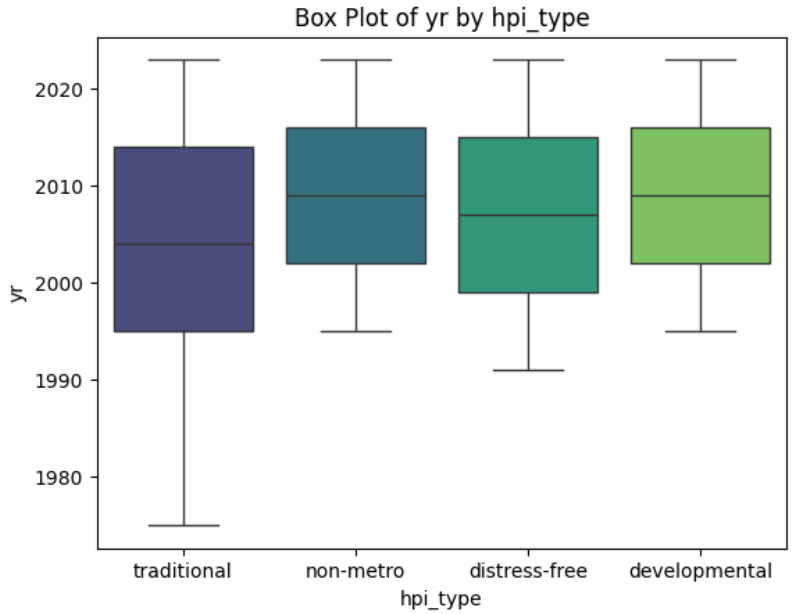
boîtes moustache en montre la distribution de l'indice NSA pour chaque catégorie de 'level'

L'ÉTUDE ENTRE 'YEAR ' ET ' HPI_TYPE '

- Le t-statistic négatif (-7,93) suggère que la valeur moyenne pour la catégorie "non-metro" est significativement inférieure à la moyenne générale de yr.

- 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.

- 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.



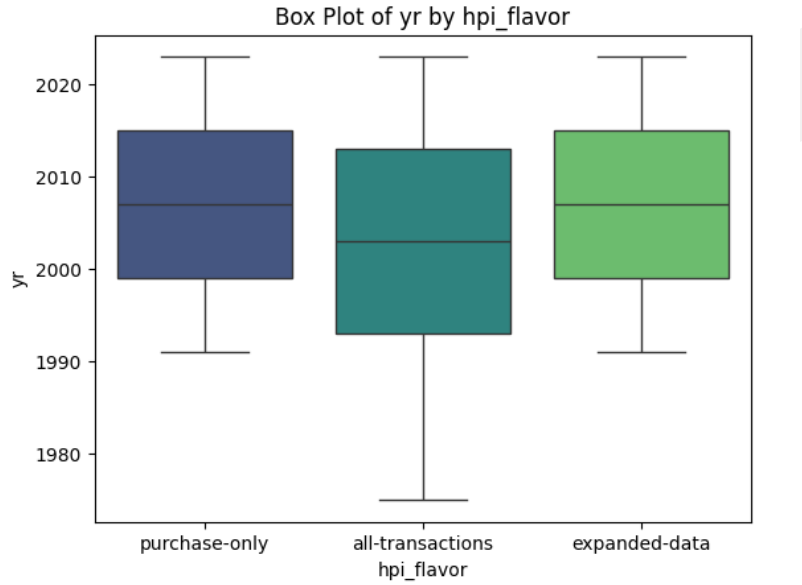
boîtes moustache en montre la distribution de l'indice year pour chaque catégorie de 'hpi_type'

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.

- 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.

- 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.



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

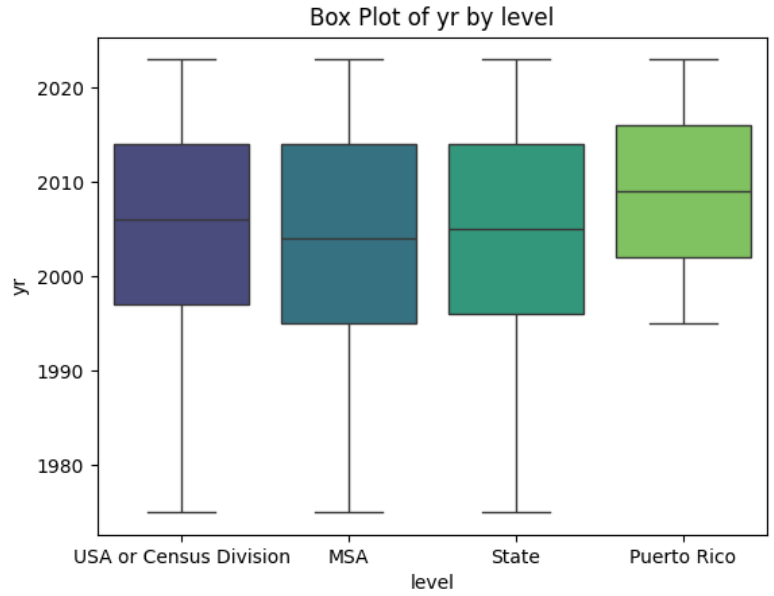
L'ÉTUDE ENTRE 'YEAR ' ET ' LEVEL '

- la moyenne globale (Moyenne: 2005.04). Cela suggère que les tendances au sein des 'USA or Census Division' ont tendance à se produire plus tard que la tendance générale.

- Pour la catégorie "MSA" est significativement antérieure à la moyenne globale (Moyenne: 2003.70),les tendances de MSA ont tendance à se produire plus tôt que la tendance globale.

- L'année moyenne (yr) pour la catégorie "State" est significativement plus tardive que la moyenne globale (Moyenne: 2004.48).La tendancede cette zone se produire plus tard que la tendance globale.

- "Puerto Rico" est significativement plus tardive que la moyenne globale (Moyenne: 2009.19).



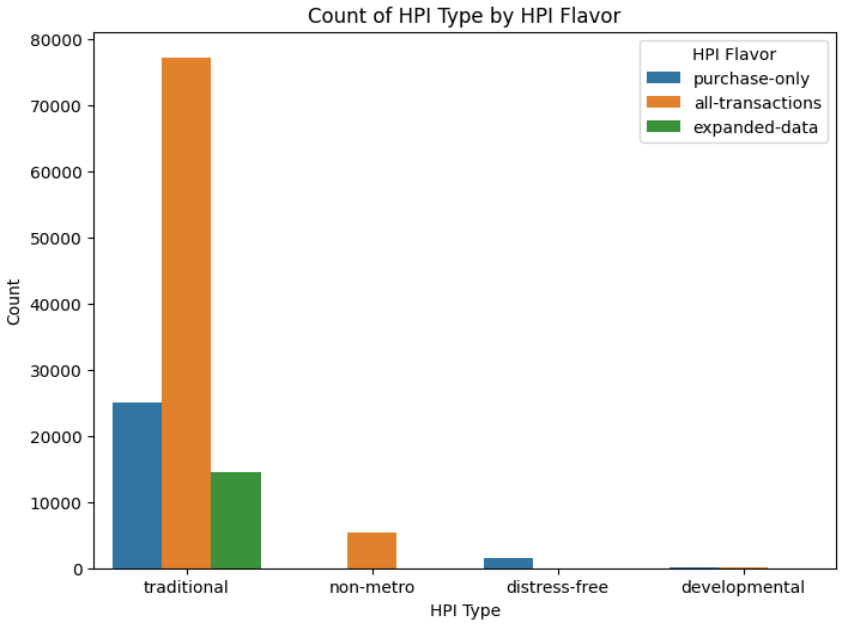
boîtes moustache en montre la distribution de l'indice year pour chaque catégorie de 'hpi_flavor'

L'ÉTUDE ENTRE ' HPI_TYPE 'ET ' HPI_FLAVOR '

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

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

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 que ces variables ne sont probablement pas indépendantes et qu'il existe une relation entre les différentes modalités de ces deux variables.



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

ANALYSE MULTIVARIÉE

L'ANALYSE MULTIVARIÉE – CHOIX DES VARIABLES :

- Parmi les dix variables disponibles dans le dataset , on s'est concentrée sur quatre variables spécifiques :

'yr' | 'period' | 'index_nsa' | 'index_sa'

- 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.

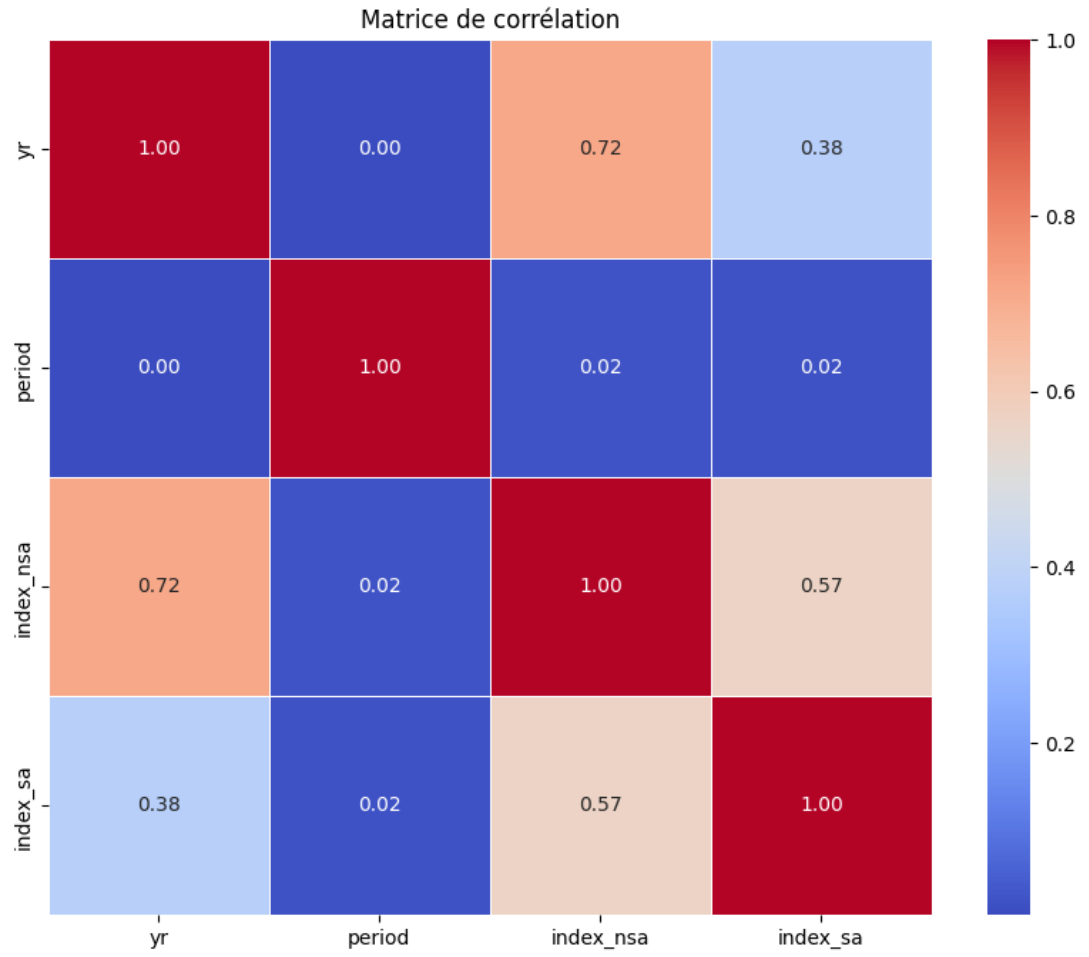
Une analyse en composantes principales (PCA) a été utilisée pour réduire la dimensionnalité des données.
Cette analyse a permis d'obtenir deux composantes principales expliquant respectivement environ 52% et 25% de la variance totale.
Les composantes principales ont été identifiées, mettant en lumière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

	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

	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

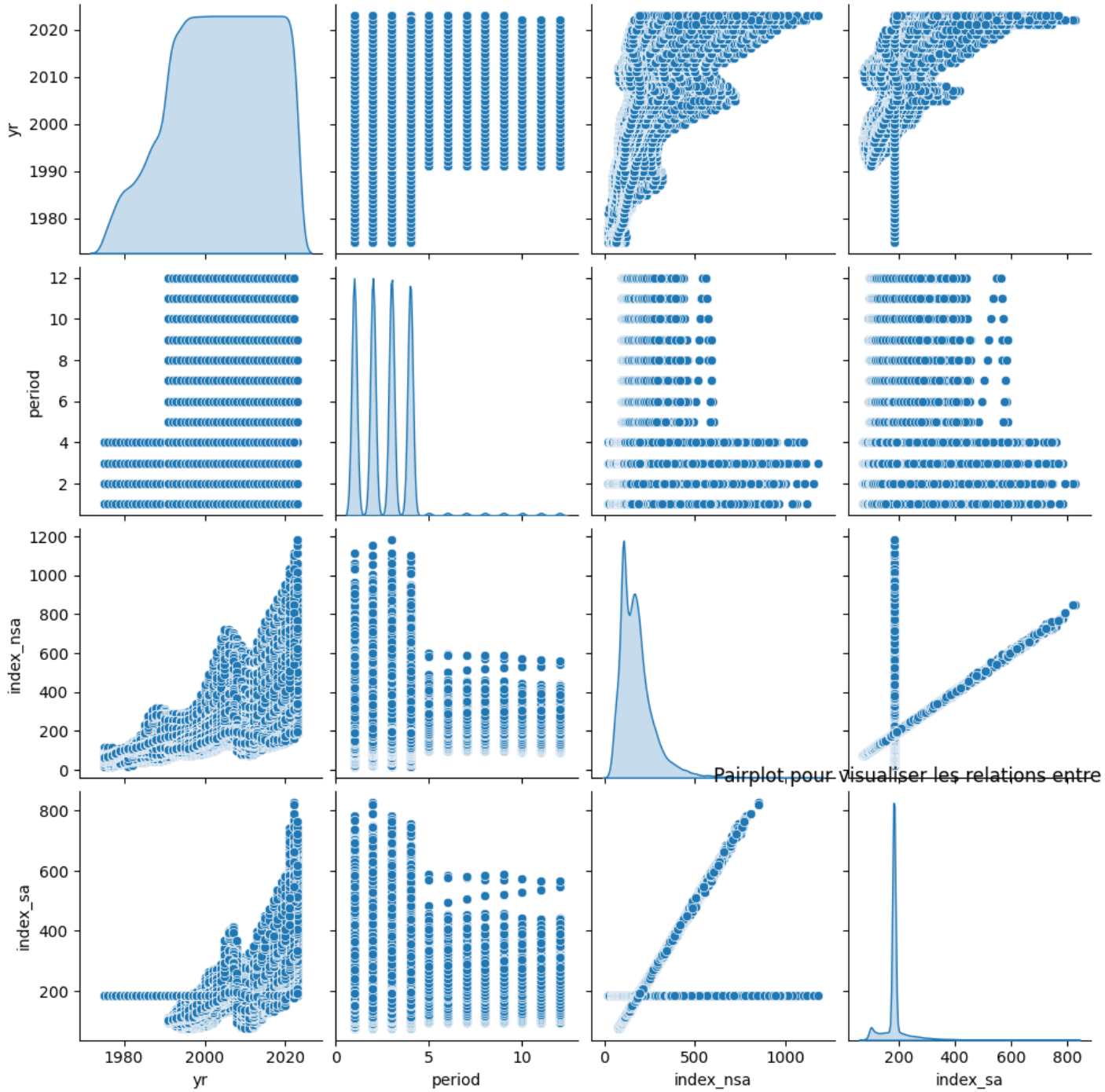
Pour visualiser l'impact de chaque composante, une représentation graphique a été réalisée en utilisant la méthode du coude (Elbow Method), montrant la décroissance des valeurs propres en fonction du nombre de composantes principales. Cette méthode permet d'identifier le nombre optimal de composantes à retenir pour expliquer au mieux la variance des données.

LA MATRICE DE CORRÉLATION

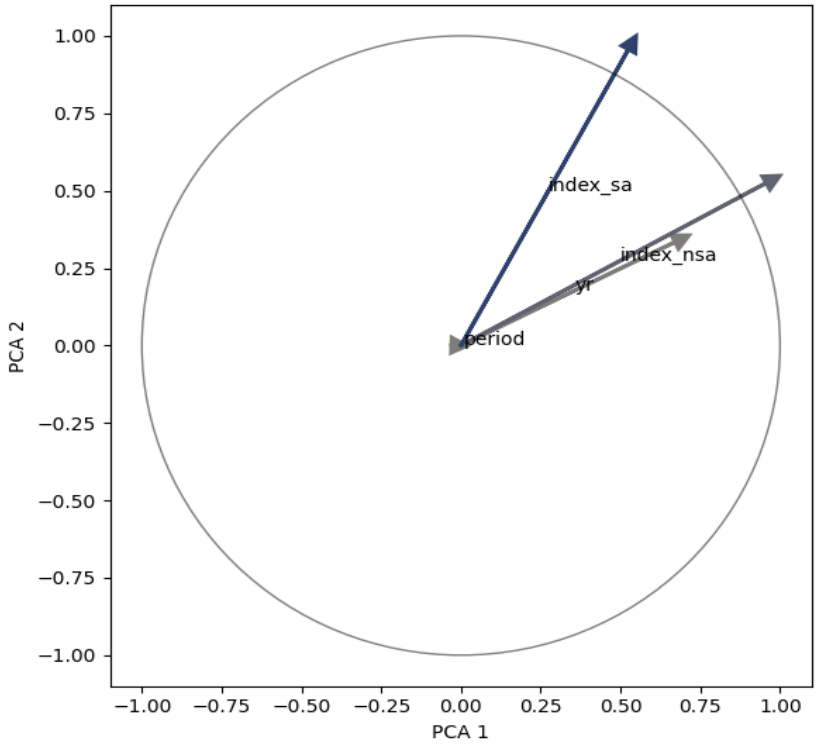


PAIRPLOT POUR VISUALISER LES RELATIONS ENTRE LES VARIABLES

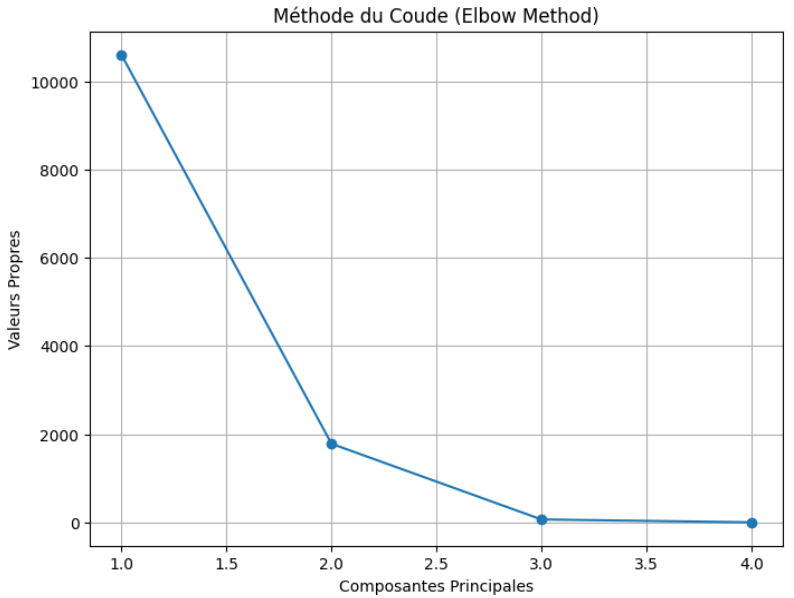
	yr	period	index_nsa	index_sa
yr	138.342387	0.069243	819.669029	669.931448
period	0.069243	2.063072	2.805549	-0.180174
index_nsa	819.669029	2.805549	9488.647942	8545.976218
index_sa	669.931448	-0.180174	8545.976218	8502.457926



Pairplot pour visualiser les relations entre les variables



CERCLE DE CORRÉLATION



METHODE DU COUDE(ELBOW METHOD)

CONCLUSION

- 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.

TEAM

HALIMA NFIDSA

Étudiante en master en Ingénierie des Systèmes Intelligents | Passionné par l'IA et l'apprentissage automatique de la machine .

NAHLA YASMINE IHOUBI

Étudiante en master en Ingénierie des Systèmes Intelligents| Passionné par l'IA et la Science des Données.

ABDELATIF MEKRI

Étudiant en master en Ingénierie des Systèmes Intelligents | Passionné par l'IA et les données .

MERCI POUR VOTRE ATTENTION

Contoso

THANK YOU!

FLORA@CONTOSO.COM

HTTP://WWW.CONTOSO.COM/