27/04/2024

# Rapport

*Analyse de données – TP2*

Marianne Corbel

# Rapport

*Analyse de données – TP2*

## Exercice 1

Cet exercice vise à fournir une analyse basique des prix de 64 paquets de pâtes alimentaires trouvés en grande surface (ici Carrefour) en fonction de leur type, marque, et poids. Tous les prix ont été relevés manuellement au mois de mars 2024.

Le fichier de données utilisé dans cet exercice est le fichier « data/pates_carrefour .csv », ressemblant à ceci :

| | Type | Marque | Poids du paquet (kg) | Prix du paquet (€) |
|---|---|---|---|---|
| **0** | Penne | Barilla | 1.0 | 1.99 |
| **1** | Penne | Barilla | 0.5 | 2.18 |
| **2** | Penne | Carrefour | 1.0 | 1.69 |
| **3** | Penne | Carrefour | 0.5 | 1.92 |
| **4** | Penne | Panzani | 1.0 | 2.19 |

### Prévalence des données

**Note :** une colonne supplémentaire est ajoutée aux données lorsqu'elles sont chargées dans le programme, le « prix au kilo ». Nous analyserons à la fois les tendances liées au prix unitaire, ainsi que celles liées au prix au kilo.

| Type | Nombre | % |
|---|---|---|
| **Coquillettes** | 10 | 15.62 |
| **Penne** | 8 | 12.50 |
| **Macaroni** | 8 | 12.50 |
| **Spaghetti** | 8 | 12.50 |
| **Farfalles** | 8 | 12.50 |
| **Fusilli** | 6 | 9.38 |
| **Tagliatelles** | 4 | 6.25 |
| **Pipe Rigate** | 4 | 6.25 |
| **Torsades** | 4 | 6.25 |
| **Lasagnes** | 3 | 4.69 |
| **Cannelloni** | 1 | 1.56 |
| | 64 | 100 |

| Marque | Nombre | % |
|---|---|---|
| **Barilla** | 18 | 28.12 |
| **Carrefour** | 16 | 25.00 |
| **Panzani** | 13 | 20.31 |
| **Rummo** | 5 | 7.81 |
| **Grand'Mère** | 4 | 6.25 |
| **Alpina Savoie** | 2 | 3.12 |
| **Lustucru** | 2 | 3.12 |
| **Granoro** | 1 | 1.56 |
| **Garofalo** | 1 | 1.56 |
| **De Cecco** | 1 | 1.56 |
| **Simpl** | 1 | 1.56 |
| | 64 | 100 |

| Poids (kg) | Nombre | % |
|---|---|---|
| **0.250** | 5 | 59 |
| **.500** | 38 | 32.81 |
| **1** | 21 | 7.81 |
| | 64 | 100 |

| Prix du paquet (€) | Nombre | % |
|---|---|---|
| **p < 1.00** | 7 | 10.93 |
| **1 <= p < 2.00** | 42 | 65.62 |
| **p > 2** | 15 | 23.43 |
| | 64 | 100 |

Grâce à ces données, nous pouvons déjà remarquer de fortes prévalences dans les données relevées en grandes surfaces :

- Poids : les paquets de 500g sont les plus nombreux, suivis par les paquets de 1kg ;
- Prix : une majorité de paquets se situent entre 1.00€ et 2.00€ par unité mais par le peu de valeurs sous 1.00€, on peut déjà s'imaginer une tendance des prix à être plus proche de 2.00€ que de 1.00€ ;
- Les marques Barilla, Carrefour et Panzani sont les plus représentées ;
- Les pâtes de type Coquillettes, Penne, Macaroni, Spaghetti et Farfalles sont les plus représentées.

Ces données peuvent avoir une importance par la suite, dans la mesure où une grande offre entraîne souvent des prix comparativement plus bas que pour un produit similaire moins représenté.

## Analyse des prix

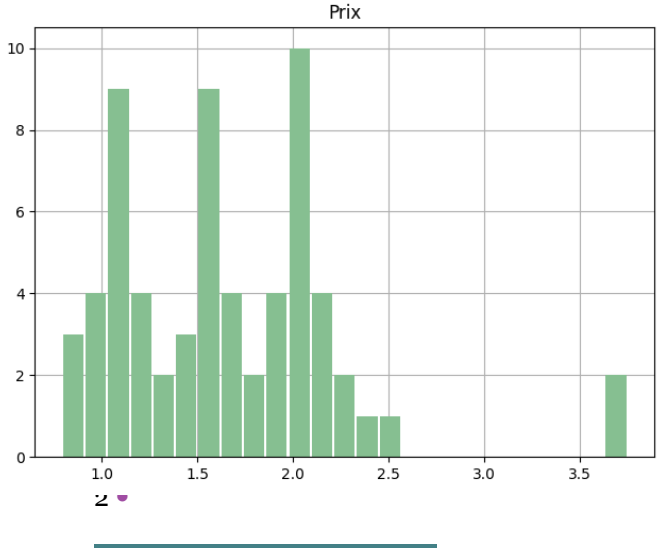
Si l'on calcule les statistiques principales du prix de la population, on en arrive aux données suivantes :

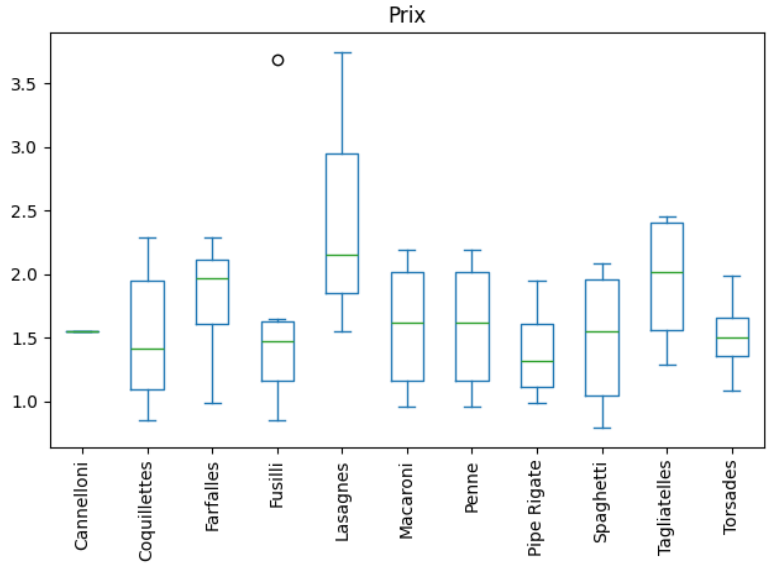
| Prix au kilo (€) | |
|---|---|
| **Moyenne** | 2.91 |
| **Médiane** | 2.19 |
| **Variance** | 2.12 |
| **Ecart-type** | 1.46 |
| **Minimal** | 0.85 |
| **25%** | 1.98 |
| **50%** | 2.19 |
| **75%** | 3.69 |
| **Maximal** | 7.50 |

| Prix du paquet (€) | |
|---|---|
| **Moyenne** | 1.65 |
| **Médiane** | 1.55 |
| **Variance** | 0.35 |
| **Ecart-type** | 0.59 |
| **Minimal** | 0.79 |
| **25%** | 1.14 |
| **50%** | 1.55 |
| **75%** | 1.99 |
| **Maximal** | 3.75 |

On peut remarquer plusieurs éléments intéressants ici :

- Les données liées au prix d'un paquet sont assez homogènes : la médiane se situe proche de la moyenne des prix, l'écart-type est moyen et la différence entre chaque quartile est relativement stable.
- Même si cela peut sembler contre-intuitif, le prix au kilo varie grandement : la moyenne est plus élevée que la médiane, indiquant des valeurs hautes moins nombreuses mais plus extrêmes. Ceci se vérifie par les valeurs du quatrième quartile : la valeur maximale est égale à plus du double de celle entre les troisième et quatrième quartile.

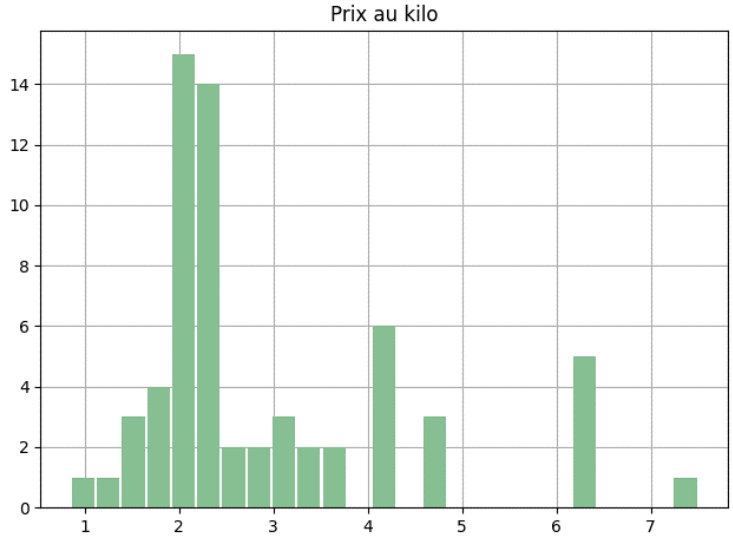


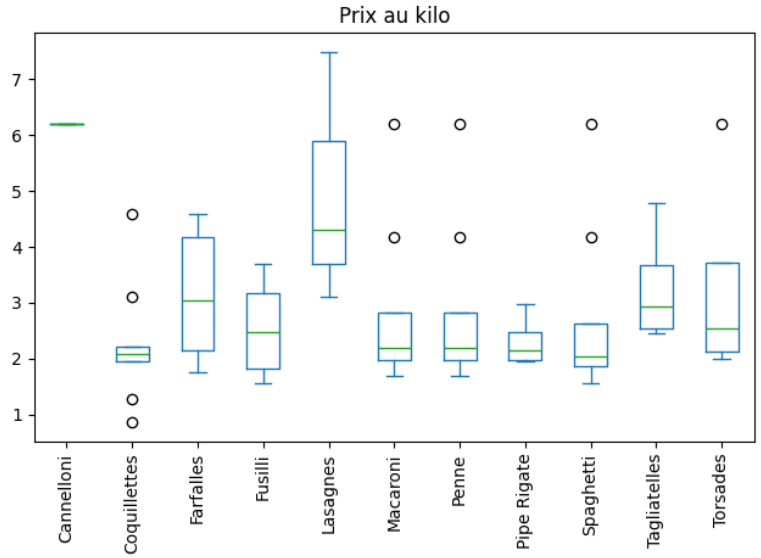



Sur les diagrammes ci-dessus, on peut de nouveau constater cette répartition des prix des paquets (indépendamment de leur poids) relativement homogène :

- La majorité des prix unitaires se situent entre 1.00€ et 2.50€, avec des pics autour de 1.00€, 1.50€ et 2.00€ (qui peuvent être interprétés comme des valeurs « pivot » entre une catégorie de prix et une autre).
- Les tranches de prix restent relativement similaires pour la majorité des types de pâtes sans grand écart entre les valeurs extrêmes et les autres. Les moyennes des prix pour tous les types de pâtes se situent entre 1.50€ et 2.00€.

On retrouve les deux valeurs extrêmes du diagramme de gauche (environ 3.70€/paquet) sur celui de droite, en tant que valeur les plus élevées des Fusilli et des Lasagnes. La valeur est considérée aberrante dans le cas des Fusillis parce qu'il s'agit typiquement d'une catégorie de pâtes à « bas prix ».




Les diagrammes ci-dessus analysent presque les mêmes données, à la différence près que le prix est au kilogramme et non à l'unité, ignorant complètement la variable de poids du paquet.

On peut constater sur le diagramme de gauche une répartition nettement moins uniforme des prix, avec presque la moitié du total des prix entre 2.00€ et 2.50€. Le diagramme de droite va dans le même sens, en soulignant cependant beaucoup plus de valeurs aberrantes, presque toutes concernant les types de pâtes les plus représentées (Coquillettes, Macaroni, Penne et Spaghettis). Sur ces types de pâtes, l'intervalle de prix entre le premier et le troisième quartile y est particulièrement bas, avec les valeurs maximales et minimales très proches de ceux-ci.

Ces prix « serrés » pourraient s'expliquer (en considérant que l'échantillon relevé dans notre grande surface est représentatif de la répartition du type de pâtes en vente à l'échelle nationale) par le fait qu'une grande offre permet moins de liberté sur les prix si l'on souhaite rester compétitif. Cette théorie peut être renforcée par le fait que les valeurs maximales des Lasagnes et des Tagliatelles, très peu représentées, sont bien plus élevées que les valeurs comprises dans leurs « boîtes » respectives.

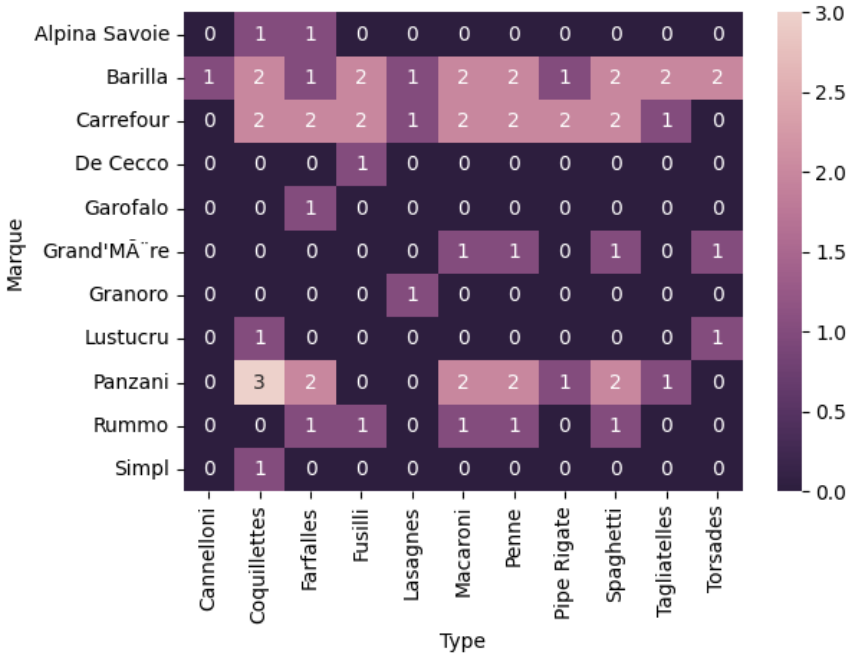
# Rapport

*Analyse de données – TP2*

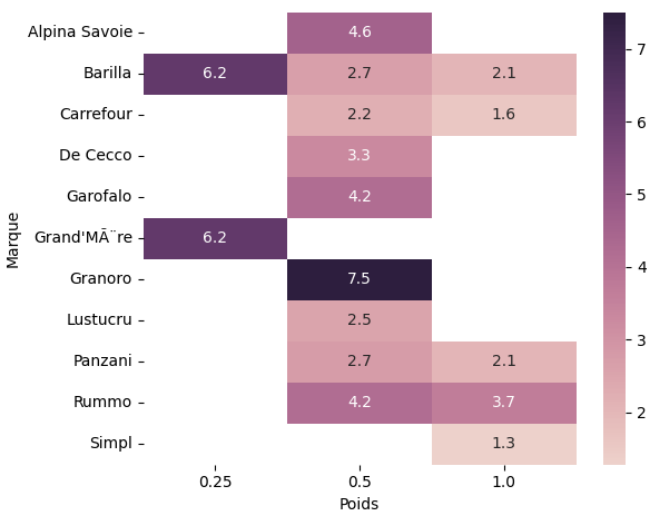
## Exercices 2 et 3

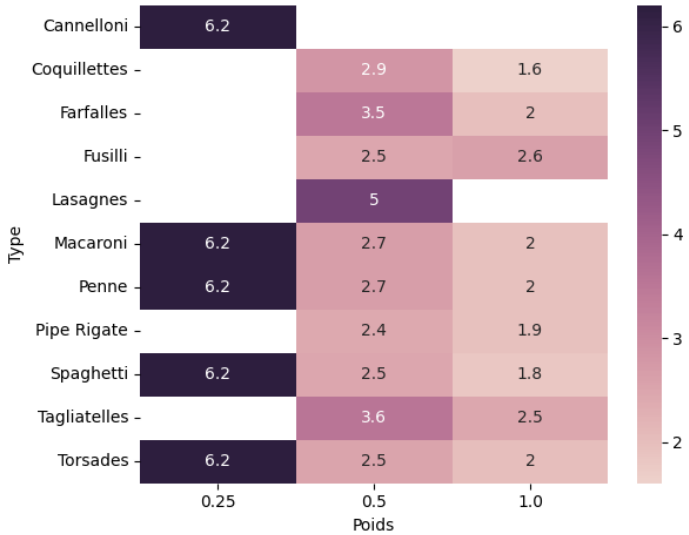
Ces exercices se situent dans la continuité de l'exercice 1 dans la mesure où ils utilisent les mêmes données, précédemment décrites dans l'introduction de celui-ci. Ils ont été regroupés par simplicité.

Le diagramme suivant décrit la répartition des types de pâtes en fonction de la marque, en nombre de paquets :




On y voit que Barilla et Carrefour et Panzani sont les marques majoritaires, liées à presque tous les types de pâtes. En dehors de ces trois marques, les types de pâtes et les marques associées sont variées. Cette répartition peut s'avérer importante pour interpréter les diagrammes suivants.



Ci-dessus, le diagramme de gauche représente la variation du prix au kilo en fonction du poids du sachet et de la marque. Plusieurs tendances sont à noter :

- Concernant les trois marques les plus représentées, on peut constater que Barilla et Panzani proposent des prix assez similaires, tandis que Carrefour place ses prix bien plus bas ;
- Les marques à faible représentation sont, à l'exception de Simpl, sensiblement plus coûteuses que les marques fortement représentées (Granoro, Grand'Mère, Garofalo). T

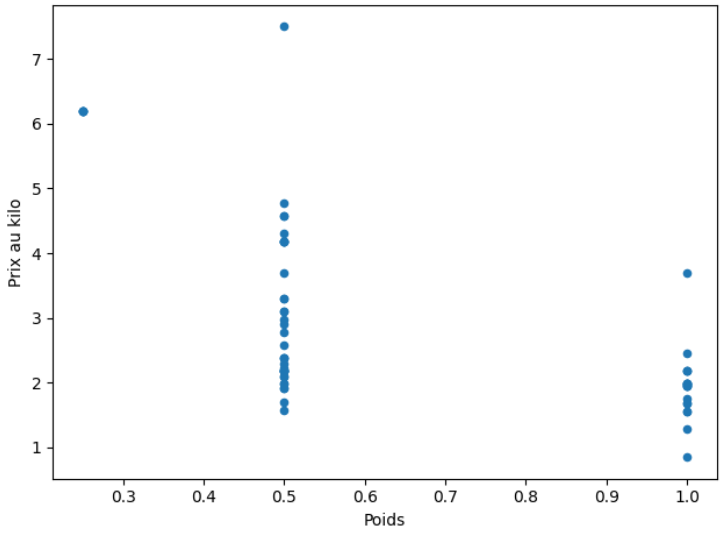
Le diagramme de droite, quant à lui, représente le prix au kilo en fonction du poids et du type de pâtes. De nouveau, plusieurs éléments sont à noter :

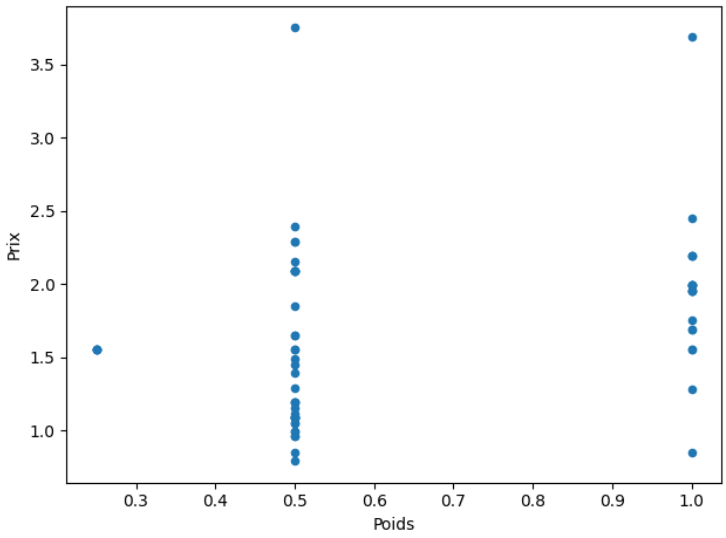
- Les types de pâtes qui apparaissaient plus cher sur les diagrammes plus hauts ressortent de nouveau comme ayant un prix moyen supérieur à la moyenne (Lasagne, Tagliatelles, Farfalles) ;
- On peut discerner une tendance du prix au kilo à baisser à mesure que le poids du sachet monte.

Cette tendance peut se vérifier à l'aide de la covariance et du coefficient de corrélation :

| **Poids / Prix au kilo** | |
|---|---|
| **Covariance** | -0.22 |
| **Corrélation** | -0.60 |

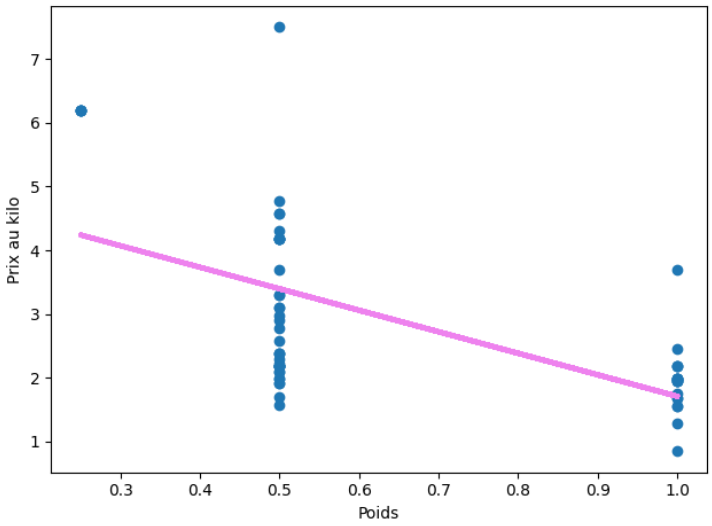
| **Poids / Prix au paquet** | |
|---|---|
| **Covariance** | 0.05 |
| **Corrélation** | 0.33 |

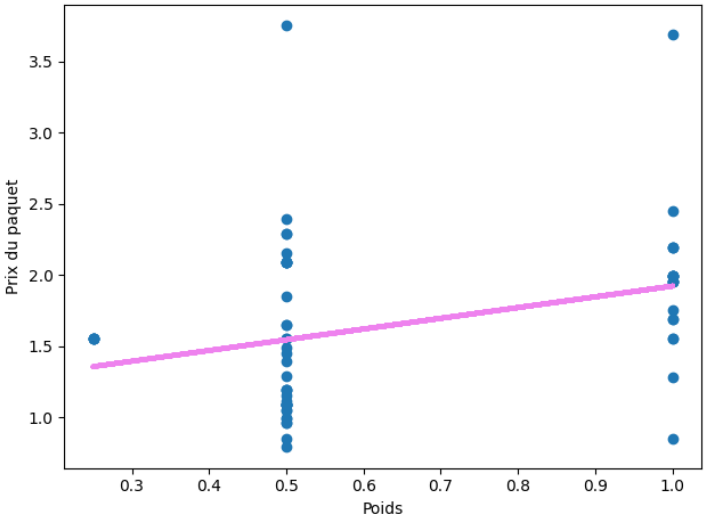



Ici, on peut voir que le prix au kilo baisse lorsque le poids du paquet augmente, comme le suggère la corrélation suffisamment significative et négative. A l'inverse, le poids du sachet augmente lorsque que le poids augmente aussi.

Les diagrammes semblent illustrer cette tendance. Pour en avoir le cœur net, on peut dessiner la droite de régression linéaire sur chacun des deux nuages de points :




Comme les deux coefficients de corrélation semblaient le suggérer, le prix, ainsi que le prix du paquet, sont donc corrélés au poids du paquet.

# Rapport

*Analyse de données – TP2*

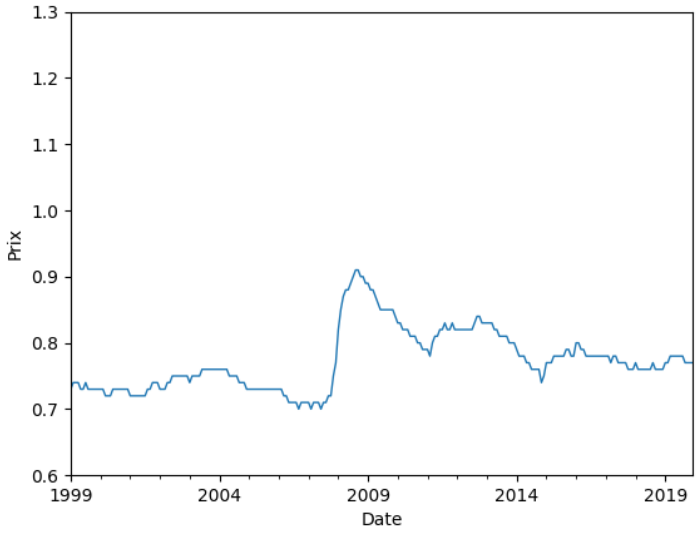
## Exercice 4

Cette analyse vise à fournir une analyse de la variation mensuelle des prix d'un sachet de pâtes alimentaires, de 2000 à 2024, en utilisant la transformée de Fourier rapide (FFT). Deux ensembles de données sont utilisés dans ce but, les deux présentant la même structure :

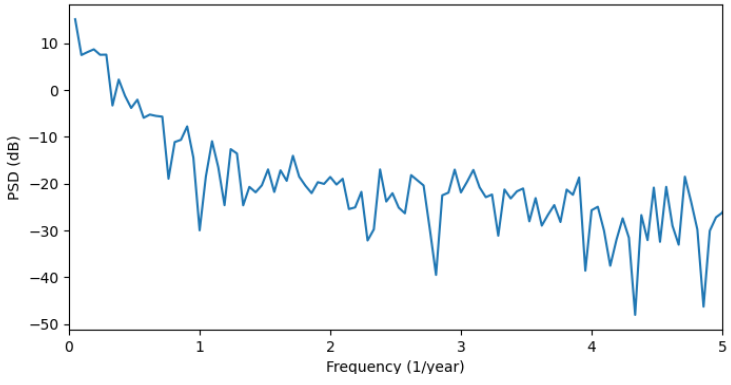
| | Date | Prix |
|---|---|---|
| **0** | 1999-01-01 | 0.73 |
| **1** | 1999-02-01 | 0.74 |
| **2** | 1999-03-01 | 0.74 |
| **3** | 1999-04-01 | 0.74 |
| **4** | 1999-05-01 | 0.73 |

### Données de l'Insee

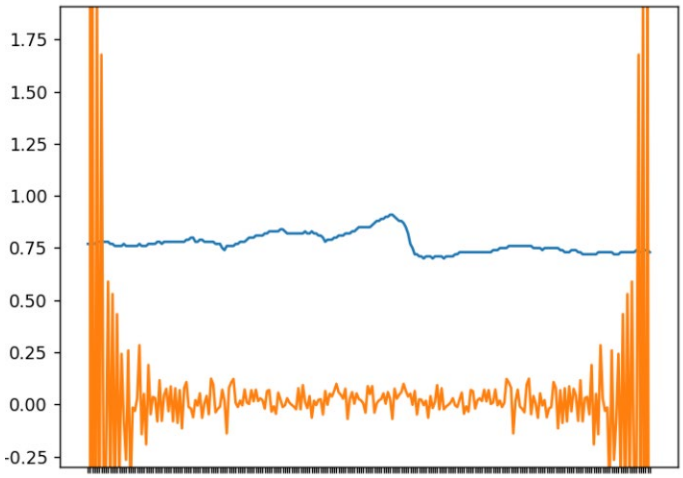
Le premier ensemble de données a été réalisé par l'Insee et contient les relevés mensuels du prix moyen d'un paquet de pâtes entre 1999 et 2019.




Ce diagramme a mené au calcul du spectre de puissance suivant :



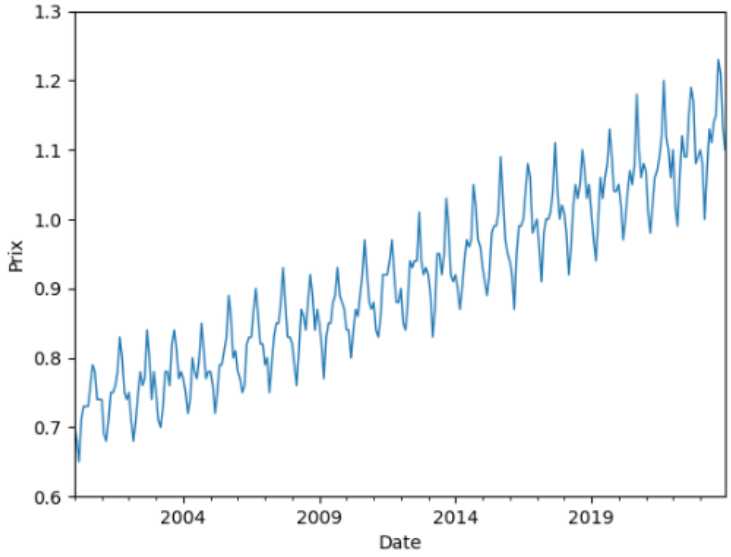
Une seconde (parmi d'autres…) tentative a révélé le spectre ci-dessous.




## Données générées

Le second ensemble contient des données générées manuellement de manière pseudo-aléatoire : le but était de simuler une forme d'inflation, tout en créant une périodicité saisonnière dans les données.

Il a été créé pour tenter de rendre la transformée de Fourier rapide plus simple, ou peut-être plus facile à comprendre, en créant artificiellement une périodicité reconnaissable.




Le même calcul a été appliqué pour obtenir le spectre de puissance suivant :