Saé S2-04

EXPLOITATION STATISTIQUE D'UNE BASE DE DONNÉES

2021-2022

Table des matières

Introduction

L'objectif de cette SAé 2.04 est de mettre en relation et d'approfondir nos connaissances en bases de données et en statistiques. La base de données utilisée pour analyser les données et en extraire les informations statistiques on a utilisé les tickets de caisse d'un magasin d'alimentation.

Pour les variables quantitatives on a utilisé l'échantillon PrixUnitaire de la population DetailTicket. Puis, pour la variable qualitative on a utilisé l'échantillon Num_Ville et Num_Ville de la population Ville et CarteFidelite . Pour finir, pour les variables quantitatives on a utilisé l'échantillon PrixUnitaire et l'unité de stock de la population Produit.

Pour tout ce qui est Excel :

Pour une variable quantitative

Pour la moyenne on a utilisé la fonction "MOYENNE" qui était accessible à partir du bouton  . Pour obtenir le résultat on a saisi : =MOYENNE(DetailTicket!C2:C24333)

Le résultat obtenu est :

moyenne		10,40040934

L'interprétation de ce résultat est que la moyenne des prix sur les 24 333 tickets est de 10,40 €.

La variance a été calculée à l'aide de la fonction "VAR" qui utilise la fonction :

$$\frac{\sum(x-\bar{x})^2}{(n-1)}$$

Ainsi, on a pris VAR et non les autres VAR.P, VAR.S ET VAR.P.N car la fonction VAR en calcule la variance en considérant que la plage Excel est un échantillon de n individus tirés dans une population plus grande.
Pour obtenir le résultat on a saisi : =VAR(DetailTicket!C2:C24333)

Le résultat obtenu est :

variance		46,26771827

En général, plus la variance est élevée, plus les données sont dispersées. Ici, nous on ne peut pas trop l'interpréter car on n'a pas d'autre variance pour l'interpréter avec.

Puis, l'écart-type a été calculé à l'aide de la fonction "ECARTYPE" qui est déjà implémentée dans Excel.
Pour obtenir le résultat on a saisi : =ECARTYPE(DetailTicket!C2:C24333)

Le résultat obtenu est :

écart-type		6,802037802

L'interprétation de ce résultat est que les prix ne sont pas trop dispersés en rapport avec la moyenne. Plus le résultat est élevé, plus la dispersion est plus importante.

La médiane a été calculé à l'aide de la fonction "MEDIANE" qui est déjà implémentée dans Excel.
Pour obtenir le résultat on a saisi : =MEDIANE(DetailTicket!C2:C24333)

Le résultat obtenu est :

médiane		8,97

L'interprétation est qu'il y'a autant de tickets en-dessus de 8.97 et en dessous.

Pour le 1er quartile et le 3e quartile on a utilisé la fonction "Quartile" qui est déjà implémentée dans Excel.
Pour obtenir le 1er quartile on a saisi : =QUARTILE(DetailTicket!C2:C24333;1)
Pour obtenir le 3e quartile on a saisi : =QUARTILE(DetailTicket!C2:C24333;3)

Le résultat obtenu est :

quartiles		
Q1		6
Q3		13,42

L'interprétation de ceci est qu'il y a au moins 25 % des prix des tickets sont inférieures ou égales à 6€ pour le 1er quartile. Puis, pour le 3e qu'il y a au moins 75% des prix des tickets sont inférieures ou égales à 13,42 € pour le 3e quartile.

Pour le 30e centiles et le 90e centiles on a utilisé la fonction "CENTILE" qui est déjà implémentée dans Excel.
Pour obtenir le 30e centile on a saisi : =CENTILE(DetailTicket!C2:C24333;0,3)
Pour obtenir le 90e centiles on a saisi : =CENTILE(DetailTicket!C2:C24333;0,9)

Le résultat obtenu est :

centiles (ou percentiles)		
30		6,53
90		16,58

Pour le tri à plat on a utilisé les fonctions "NB.SI" et "NB.SI.ENS". Pour la 1ere la requête j'ai utilisé un NB.SI seulement car j'ai utilisé qu'une seule condition car nous ne pouvons pas avoir un prix inférieur à 0 donc j'ai seulement besoin d'une seule condition qui est une valeur inférieure ou égale à 5
Alors que dans les 2 autres requêtes il me faut 2 conditions en s'assurant par exemple dans la deuxième requête que la valeur soit strictement supérieure à 5 et dans la troisième requête que la valeur soit strictement supérieure à 10.

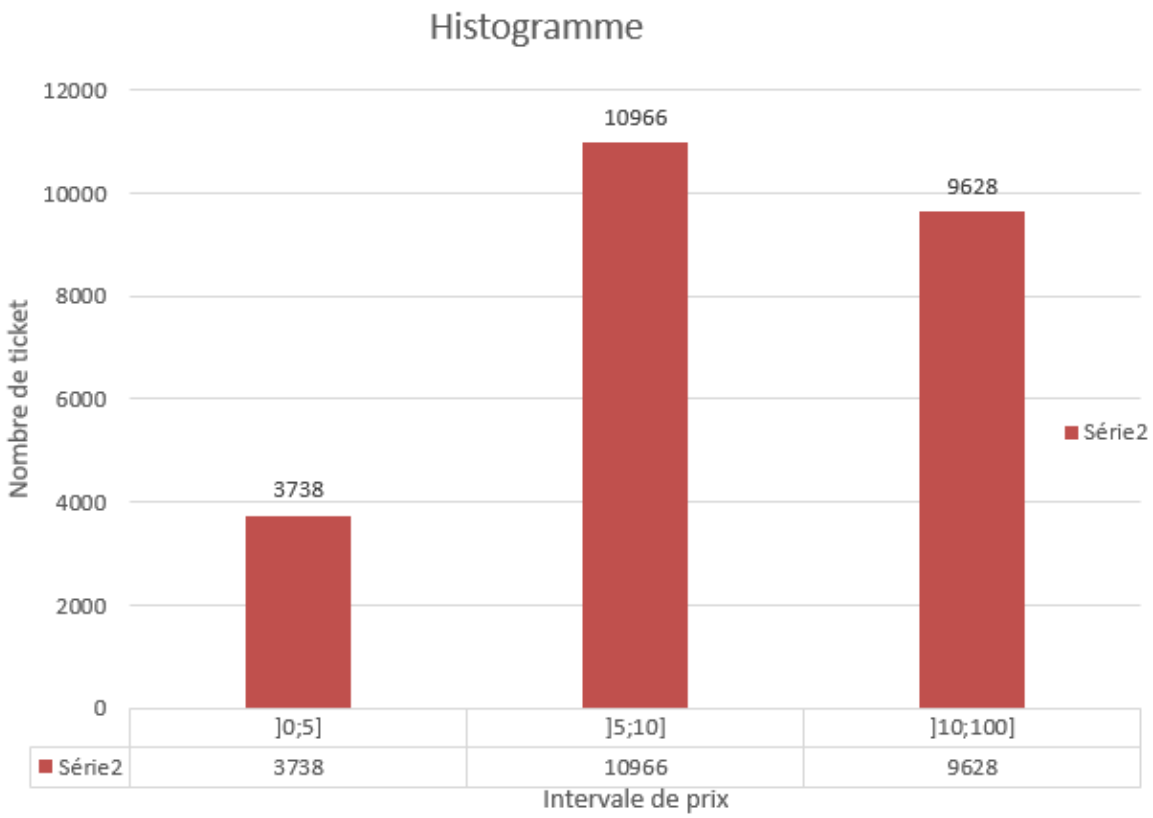
Pour obtenir le tri à plat de]0 ;5] : =NB.SI(DetailTicket!C2:C24333;"<=5")
Pour obtenir le tri à plat de]5 ;10] :
=NB.SI.ENS(DetailTicket!C2:C24333;"<=10";DetailTicket!C2:C24333;">5")
Pour obtenir le tri à plat de]10 ;100] :
=NB.SI.ENS(DetailTicket!C2:C24333;"<=100";DetailTicket!C2:C24333;">10")

Le résultat obtenu est :

tri à plat		
]0;5]		3738
]5;10]		10966
]10;100]		9628

Pour le tri à plat j'ai d'abord dans un premier temps répartis les prix des tickets, j'ai choisi tous les prix du tickets qui valent entre 0 et 5 euros, ensuite nous avons ceux qui valent entre 5 et 10 euros et pour terminer ceux qui valent entre 10 et 100 euros
Pour réaliser le tri à plat j'ai seulement sélectionné parmi les tickets tous les prix qui valent entre 0 et 5 euros, c'est à dire les prix inférieur ou égale à 5 euros (parce que nous n'avons pas de prix négatif).
Ensuite pour les prix compris entre 5 et 10 j'ai sélectionné tous les prix qui sont strictement supérieur à 5 et inférieur ou égale à 10 euros.
Et pour terminer pour les prix compris entre 10 et 100 euros j'ai sélectionné tous les prix qui sont strictement supérieur à 10 et inférieur ou égale à 100.
Pour conclure, ces requêtes permettent de déterminer le nombre de tickets en fonction d'un intervalle de prix.

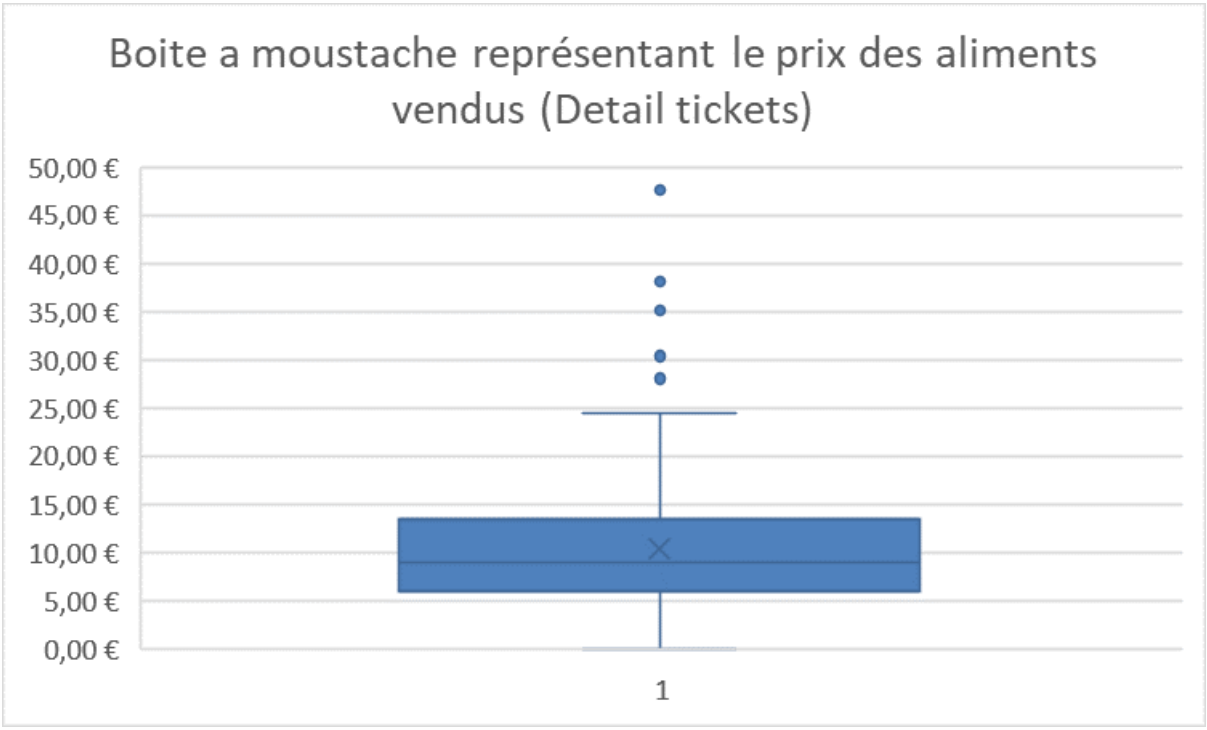
L'histogramme qui représente le tri à plat.



Nous avons un histogramme qui représente le nombre de tickets compris entre chaque intervalle de prix.

Nous avons une boite à moustache qui résume les indicateurs de notre liste de prix des tickets.
Nous pouvons voir le minimum qui est à 0 euros, le premier quartile qui est bien à 6 euros (représenté par le début du carré de la boite) et un troisième quartile (représenter à la fin du carré de la boîte) à 13 euros. Nous pouvons remarquer que ces valeurs peuvent être confirmées avec les valeurs que nous avons calculées précédemment.

BOITE A MOUSTACHE STRATIFIÉ :



J’ai tout d’abord séparer deux éléments pour créer notre boîte à moustache stratifié:

D’un côté les aliments et de l’autre les vêtements

Rayon Alimentaire	3086
Rayon Entretien et Hygien	0
Rayon Vêtement	29
Rayon Divers	0
Rayon Animaux	0

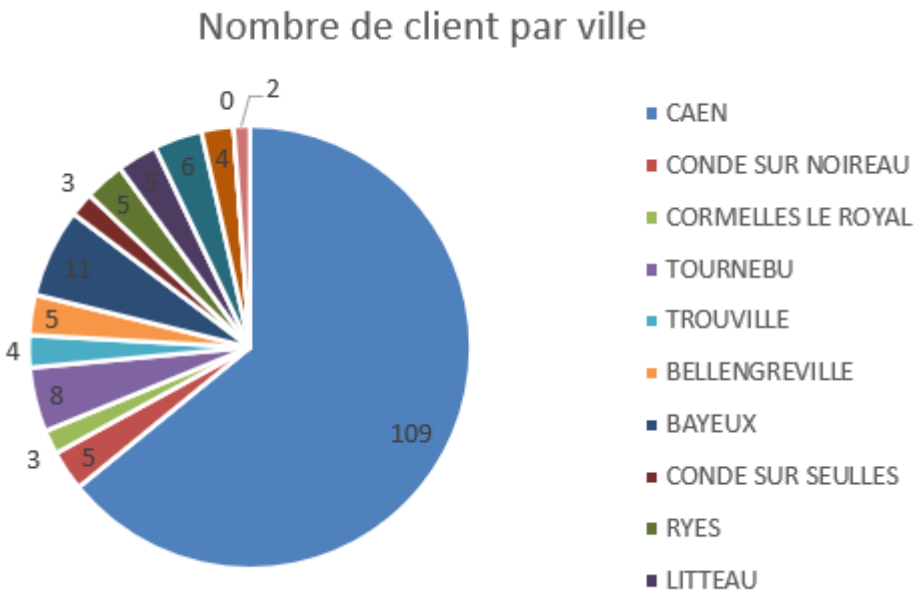
Ensuite dans la liste des détails tickets j’ai rassembler tous les aliments ensemble et les vêtements d’un côté. J’ai pu constater que la grande majorité des produits vendus était des aliments et seulement 8 produits vendus étaient des vêtements.

362	49450	6,34 €	4	0,00%
362	48319	6,34 €	4	0,00%
362	47879	6,34 €	4	0,00%
362	47042	6,34 €	4	0,00%
362	24450	6,34 €	3	0,00%
362	23319	6,34 €	3	0,00%
362	22879	6,34 €	3	0,00%
362	22042	6,34 €	3	0,00%

Mais malheureusement représenter les vêtements sous forme d’une boîte à moustache n'était pas possible car nous n'avons qu’un seul prix à chaque fois.

Pour une variable qualitative

Pour la variable qualitative on essaye d'analyser le nombre de personnes dans chaque ville à l’aide d’un camembert.



CAEN	109
CONDE SUR NOIREAU	5
CORMELLES LE ROYAL	3
TOURNEBU	8
TROUVILLE	4
BELLENGREVILLE	5
BAYEUX	11
CONDE SUR SEULLES	3
RYES	5
LITTEAU	5
GRENTHEVILLE	6
BLONVILLE SUR MER	4
HEROUVLLE SAINT CLAIR	0
MONDEVILLE	2

Axe Horizontal (Catégorie)

Pour réaliser ce tableau nous avons d'abord trier tous les num_ville des clients en ordre croissant pour nous faciliter à sélectionner toutes les personnes de la ville numéro 1 par exemple et pour terminer nous avons rassemblé toutes les personnes d'une ville ce qui nous a permis de créer un camembert représentant le nombre de client par ville

Pour deux variables quantitatives

COVARIANCE

covariance		-22,08963772

Pour réaliser la covariance de deux variables quantitatives, j'ai d'abord sélectionné deux variables qui sont les Unités de stocks et les Unités de commande
La covariance permettrait de quantifier les écarts conjoints des deux variables par rapport à leurs espérances respectives.
Pour réaliser la covariance des deux variables j'ai réalisé la requête qui est la suivante.

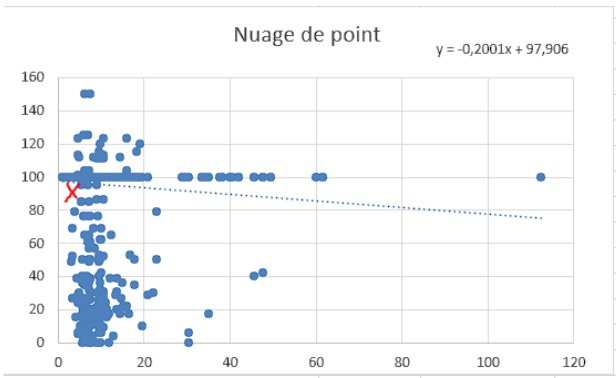
```
=COVARIANCE(Produit!I2:I3116;Produit!H2:H3116)
```

COEFFICIENT DE CORRÉLATION

coefficient de corrélation		-0,325874303

Pour le coefficient de corrélation nous avons repris les mêmes variables quantitatives que celle précédemment. Le coefficient de corrélation est la mesure spécifique qui quantifie la force de la relation linéaire entre les deux variables.
Pour calculer le coefficient de corrélation des deux variables j'ai utilisé la requête qui est la suivante.

```
=COEFFICIENT.CORRELATION(Produit!H2:H3116;Produit!I2:I3116)
```



Rayon Vêtement
Rayon Divers
Rayon Animaux

Moyenne Unité DE STOCK	Moyenne Unité de commande
96,219904	0,2616372

Le centre de gravité se trouve au point, (0,2616372;96,219904)

x : centre de gravité.

Pour générer le nuage de points j'ai sélectionné dans un premier temps les deux variables quantitatives qui sont toujours les unités de stocks et les unités de commande ensuite j'ai inséré un nuage de points dont je rajouterais à la suite une droite de régression et son équation qui se trouve au-dessus du nuage de point.
Pour le centre de gravité j'ai eu une difficulté à trouver comment générer le centre de gravité, je l'ai alors créé moi-même en calculant la moyenne des deux variables ce qui nous donne les coordonnées du centre de gravité.

Pour tout ce qui est python :

Pour un ensemble d'au moins quatre variables quantitatives

Pour l'ensemble d'au moins quatre variables quantitatives on a utilisé l'échantillon PrixUnitaire, Unites_Stock, Unites_Com et Niveau_Reap de la population Produit.

Pour la matrice de corrélation on a écrit cette commande :

```
v = ["Prix_unit" , "Unites_Stock" , "Unites_Com" , "Niveau_Reap"]
d = df[ variables ].values
E = ((d - mean(d, axis = 0)) / std(d, axis = 0))
Result = 1 / len(df ["Prix Unit"]) * dot(E.T , E)
print("La matrice de corrélation est\n", Result)
```

Pour avoir ce résultat :

```
La matrice de corrélation est

[[ 1.         -0.05944259  0.03455501  0.08762233]
 [-0.05944259  1.         -0.3258743  -0.35225225]
 [ 0.03455501 -0.3258743   1.          0.51881209]
 [ 0.08762233 -0.35225225  0.51881209  1.        ]]
```

Puis, pour la paire de variables les plus corrélées positivement et la paire de variables les plus corrélées négativement on a écrit cette commande :

```
O = Result - eye(len(v), len(v))

P1 = v[argmin(O) // len(v)]
P2 = v[argmin(O) % len(v)]
print("La paire de variable la plus corrélée négativement est", P1,
"et", P2, "avec un coefficient de corrélation de", O[argmin(O) //
len(v)] [argmin(O) % len(v)])

P3 = v[argmax(O) // len(v)]
P4 = v[argmax(O) % len(v)]
print("\nLa paire de variable la plus corrélée positivement est", P3,
"et", P4, "avec un coefficient de corrélation de", O[argmax(O) //
len(v)] [argmax(O) % len(v)])
```

Pour avoir ce résultat :

```
La paire de variable la plus corrélée négativement est Unites_Stock et
Niveau_Reap avec un coefficient de corrélation de -0,35225225415617584

La paire de variable la plus corrélée positivement est Unites_Com et
Niveau_Reap avec un coefficient de corrélation de
0.5188120860899116
```

Pour une variable quantitative

Pour les variables quantitatives on a utilisé l'échantillon PrixUnitaire de la population DetailTicket.

Pour la moyenne on a écrit cette commande :

```
Moyenne = (sum(d)/len(d))
print("La moyenne est : ", Moyenne)
```

Pour avoir ce résultat :

```
La moyenne est :  10.406855687047996
```

Pour la variance on a écrit cette commande :

```
Variance = np.var(d)
print("La varaince est : ",Variance)
```

Pour avoir ce résultat :

```
La varaince est :  46.51099387726715
```

Pour l'écart-type on a écrit cette commande :

```
Ecart_Type = std(d)
print("L'ecart-type est : ",Ecart_Type)
```

Pour avoir ce résultat :

```
L'ecart-type est :  6.819896911043974
```

Pour la médiane on a écrit cette commande :

```
d.sort()
Mediane = median(d)
print("La mediane est : ",Mediane)
```

Pour avoir ce résultat :

```
La mediane est :  8.97
```

Pour le 1er quartile et le 3e on a écrit cette commande :

```
#1er quartile
Quartile_1 = round(np.percentile(d, 25))

#3e quartile
Quartile_3 = round(np.percentile(d, 75))
print("1er quartile : ", Quartile_1)
print("\n3e quartile : ", Quartile_3)
```

Pour avoir ce résultat :

```
1er quartile :  6

3e quartile :  13
```

Pour le 1er quartile et le 3e on a écrit cette commande :

```
Centiles_30 = np.percentile(d, 30)
Centiles_90 = np.percentile(d, 90)
print("Le 30e cartile est : ", Centiles_30)
print("Le 90e cartile est : ",Centiles_90)
```

Pour avoir ce résultat :

```
Le 30e cartile est :  6.53
Le 90e cartile est :  16.58
```

Le tableau de synthèse

Récapitulatif des notions abordées	Notions traitées dans la Saé	
Pour une variable quantitative	**Python**	**Excel**
moyenne		
variance		
écart-type		
médiane		
quartiles		
centiles (ou percentiles)		
tri à plat		
histogramme (graphique)		
boîte à moustaches (graphique)		
boîte à moustaches sur population stratifiée		
Pour une variable qualitative		
mode		
camembert/histogramme (graphique)		
Pour deux variables quantitatives		
covariance		
coefficient de corrélation		
nuage de points (graphique)		
centre de gravité (graphique)		
droite de régression (équation et tracé)		
Pour un ensemble d'au moins quatre variables quantitatives		
matrice de corrélation		
paire de variables les plus corrélées positivement		
paire de variables les plus corrélées négativement		