Examen
Data Exploration - ING1 Mathématiques Appliquées

2 feuilles R/V manuscrites autorisées, Calculatrice autorisée	*Date :* **17 décembre 2024** *Durée :* **2h** *Nombre de pages :* **4**

Il sera tenu compte de la qualité de la rédaction et de la précision des justifications.

◇ ◇ ◇

Le jeu de données *Avocado Prices* (ou Prix des Avocats) contient des informations sur les prix des avocats vendus aux États-Unis sur une période de 130 jours, ainsi que d'autres caractéristiques liées à leur vente. Nous nous intéressons particulièrement aux variables quantitatives et aux variables qualitatives suivantes :

- Price : Le prix moyen d'un kg d'avocat en dollars.
- Volume : Le volume total d'avocats vendus (en tonnes).
- Hass : Le volume d'avocats de variété Hass vendus (en tonnes).
- Fuerte : Le volume d'avocats de variété Fuerte vendus (en tonnes).
- Bacon : Le volume d'avocats de variété Bacon vendus (en tonnes).
- Year : L'année de la vente de l'avocat.
- Total Bags : Le nombre total de sacs vendus.
- Type : Le type d'avocat, généralement soit «conventional» ou «organic».
- Region : La région géographique aux États-Unis où l'avocat a été vendu « Center», «North», «South», « West».

A Statistiques univariées

On commence par étudier la variable *Price*. On obtient avec le logiciel R les résultats suivants :

```
> summary(Dataset$Price)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
2.600   3.930   4.580   4.723   5.200   7.760
```

1. Donner et interpréter les quartiles.
2. Calculer l'étendue et l'écart interquartile.
3. Calculer les moustaches m et M du boxplot.
4. Y a-t-il des valeurs atypiques pour cette variable ? Justifier.
5. Comparer la moyenne et la médiane de la variable *Price*.

B Statistiques Bivariées

B - 1 Analyse Quantitative x Quantitative

On s'intéresse au lien entre les variables $X = Price$ et $Y = Volume$. On donne les résultats suivants :

$$\bar{Y} = 522.7787, \quad \bar{X} = 4.723, \quad s_X^2 = 1.151, \quad s_Y^2 = 423428, \quad \mathrm{Cov}(X, Y) = 324.6$$

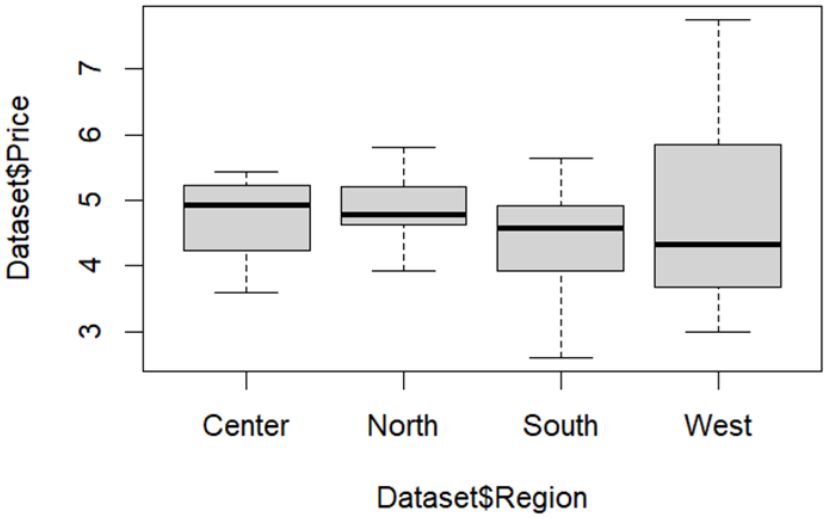
1. Calculer et interpréter le coefficient de corrélation linéaire entre X et Y.
2. Déterminer l'équation de la droite de régression de Y en fonction de X.
3. Pour un prix $X = 3.92$, la volume de vente était de 27.357 tonnes.
   (a) Donner la valeur de Y estimé pour $X = 3.92$.
   (b) Calculer l'erreur de l'estimation.
4. Calculer et interpréter le coefficient de détermination R^2.
   On donne les résumés numériques de la variable des résidus centrés :

```
> summary(rstandard(Regression))
    Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-1.3518229 -0.8825172 -0.3847021  0.0008227  0.9650227  2.5725342
```

5. Y a-t-il des valeurs atypiques ? Justifier.
6. Que pensez-vous de ce modèle ? Peut-on l'améliorer ? Justifier.

B - 2 Analyse Quantitative x Qualitative

On s'intéresse maintenant au lien entre les variables *Price* et *Region*.



1. Y a-t-il une différence de prix des avocats entre les régions ? Justifiez votre réponse en vous basant sur le boxplot.

```
> Resultat<- lm (Dataset$Price~as.factor(Dataset$Region))
> anova(Resultat)
Analysis of Variance Table
Response: Dataset$Price
                            Df  Sum Sq Mean Sq F value Pr(>F)
as.factor(Dataset$Region)    3   2.383  0.79423   0.685 0.5628
Residuals                      126 146.101   1.15953
```

2. Calculer les variances inter-groupes, intra-groupes et la variance totale.
3. Calculer et interpréter le rapport de corrélation entre les variables.

B - 3 Analyse Qualitative x Qualitative

Le tableau de contingence suivant donne la répartition des avocats selon leur type ainsi que leur région de vente.

	Center	North	South	West
Conventional	8	20	22	34
Organic	4	4	7	31

1. Donner les effectifs marginaux des deux variables Type et Région.
2. Calculer et interpréter les valeurs suivantes : f_{12}, $f_{\cdot 1}$, $f_{2|1}$.
3. Donner le profil moyen ligne (profil marginal ligne)
4. Etude des profils lignes.
   (a) Donner le tableau des profils lignes.
   (b) Comparer le tableau des profils lignes avec le profil moyen ligne.
   (c) Que pouvez-vous en conclure sur le lien entre les deux variables ? Justifier.
5. Afin de valider la conclusion de la partie précédente sur le lien entre le type des avocats et la région géographique où les avocats ont été vendus, on réalise un test statistique spécifique :
   (a) Quel est le nom du test à effectuer pour étudier le lien ?
   (b) Donner le tableau des effectifs théoriques.
   (c) La distance de khi deux est de 9.6106. Les variables sont-elles liées ?
   (d) Faites-vous confiance à votre test ? Justifier.

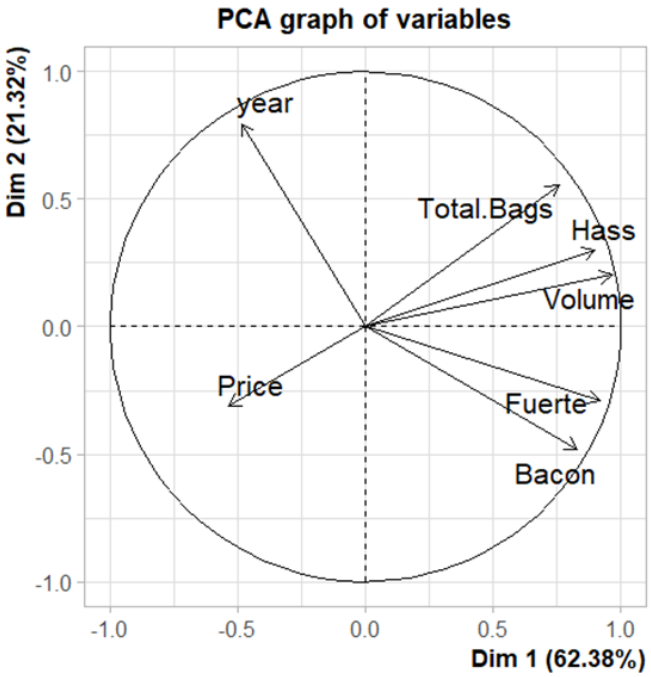
d.d.l	1	2	3	4	5	6	7	8	9	10
Seuil	3.84	5.99	7.82	9.49	11.075	12.59	14.07	15.51	16.92	18.31

C Statistiques multivariées - Analyse en Composantes Principales

On cherche maintenant à étudier le lien entre toutes les variables quantitatives à l'aide d'une analyse en composantes principales normalisée. En appliquant cette méthode avec R, les résultats sont donnés à la fin de cet exercice.

```
> ACP$eig
         eigenvalue      %of variance                cumulative % of variance
comp 1  4.366933          6.238476e+01                          62.38476
comp 2  1.492401          2.132001e+01                          83.70478
comp 3  ********          1.035451e+01                          94.05928
comp 4  ********          3.676847e+00                          97.73613
comp 5  0.1192418         1.703455e+00                          99.43959
comp 6  0.03922894        5.604134e-01                          100.00000
comp 7  1.221348e-13      1.744783e-12                          100.00000

> ACP$var$cos2
                Dim.1        Dim.2       Dim.3        Dim.4
Price         0.2871584 0.09814368 0.6065980400 8.075501e-03
Volume        0.9346742 0.04145943 0.0226223177 1.315876e-05
Hass          0.8041365 0.08876011 0.0350229083 1.224349e-02
Fuerte        0.8440085 0.08467719 0.0004123339 4.828304e-02
Bacon         0.6894436 0.23385914 0.0056258958 5.220754e-02
Total.Bags    0.5741803 0.31016398 0.0542149242 7.925708e-03
year          0.2333320 0.63533737 0.0003191980 1.286309e-01
```



1. Combien y-a-t-il de points dans le nuage de points ? En quelle dimension sont représentés ces points ?
2. Comment est définie la quantité d'information contenue dans le nuage de points ? A partir de quelle matrice peut-on la calculer et comment la calcule-t-on ?
3. Donner les deux valeurs propres manquantes dans le tableau en justifiant vos calculs.
4. Quel pourcentage de l'inertie totale contiennent les deux premières composantes principales ?
5. Donner le poids en moyenne de chaque variable ainsi que de chaque observation.
6. Toutes les variables sont-elles bien représentées en dimension 2. Justifier.
7. Interpréter la liaison entre les variables.
8. Expliquer sans faire de calculs, comment les nouveaux axes (composantes principales C_1 et C_2) sont calculés ?