CY TECH	ING1-GM EXAMEN DE DATA EXPLORATION 2020-2021
2h	Examen papier 2 feuilles R/V manuscrites autorisées Calculatrice autorisée
Nom :	Prénom :

total : 23 1/2

Groupe :

Exercice 1 — 8 pts

Les statistiques sur l'opinion politique selon l'ethnie d'un bureau de vote sont données dans le tableau suivant (données imaginaires) :

	Democrat	Independent	Republican	Sum
Asian	13	7	20	40
Black	65	18	64	147
Hispanic	107	30	64	201
White	280	90	210	580
Other	5	7	20	32
Sum	470	152	378	1000

1) Donner la valeur et la signification de n_{23}; $n_{2\bullet}$; $f_{\bullet 3}$; $f_{5|1}$ — (2)

	Valeur	Interprétation
n_{23}	64	Black, republican
$n_{2\bullet}$	147	eff. - Black
$f_{\bullet 3}$	$\frac{378}{1000}$	fréq de republican
$f_{5|1}$	$\frac{5}{470}$	fréq qu'il est "other" sachant qu'il est democrat ou 1% des Democrat sont "other".

2) Quel est le profil moyen des lignes ? — (1/2)

profil moyen des lignes :
0,47 0,152 0,378

3) Déterminer tous les profils lignes. — (1)

			total
0.325	0.175	0,5	1
0,44	0,122	0,435	1
0,532	0,149	0,318	1
0,482	0,155	0,362	1
0,156	0,218	0,625	1

4) Est-ce qu'il y a des modalités de variable qui s'écartent du profil moyen ? Que pouvez-vous en conclure ? — (1)

- tous les modalités s'écartent du profil moyen. Alors il semble avoir de lien entre les modalités de ces deux variables.

Avec le logiciel R, nous obtenons le tableau du chi-deux suivant :

```
>res=chisq.test(data)
> res$expected
          democrat  independent  republican
asian        18.80        6.080      15.120
black        69.09       22.344      55.566
hispanic     94.47       30.552      75.978
white       272.60       88.160     219.240
other        15.04        4.864      12.096
```

5) Expliquer comment est obtenu la valeur 18.80 de la case democrat/asian.

$\frac{40 \times 470}{1000}$ [effec asian × eff democrat] / eff total (1)

ce sont les eff theoriques.

6) Expliquer comment calculer la distance du chi-deux à partir du tableau. On admet que cette distance vaut 22,864.

$$X^2 = \sum_i \sum_j \frac{(\text{eff obs} - \text{eff theo})^2}{\text{eff theo}}$$ (1)

Voici le tableau de seuils de décision

d.d.l.	5	6	7	8	10	15
Seuil	11.075	12.59	14.07	15.51	18.31	24.99

7) Donner les *d.d.l.* dans le cas de ce jeu de données.

$d.d.\ell = (5-1) \times (3-1) = (4) \times (2) = 8$ (1/2)

8) Que pouvez-vous conclure sur la relation entre ces deux variables.

d.d.ℓ = 8 abs Seuil = 15.51 = C on a : $X^2 = 22,864$

$X^2 > C \Rightarrow$ les deux vaibl sont dépendants. (1)

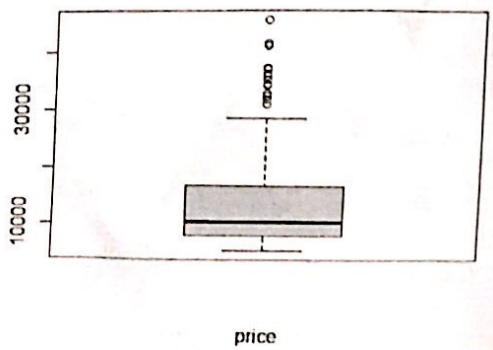
Exercice 2 (8 pts)

Pour cet exercice, nous allons utiliser le jeu de données cars.csv que vous aviez à travailler en autonomie. On rappelle que le jeu de données est constitué de 193 voitures caractérisées par 25 variables quantitatives et qualitatives.

1) Etude de la variable prix (en dollars)

Le logiciel R, nous fournit les résultats suivants :

```
>summary(cars$price)
>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5118    7738   10245   13285   16515   45400
```



1.a) Donner les quartiles et interpréter les.

$Q_1 = 7738$ → 25% des voitures ont un prix inf à 7738 $

$Q_2 = \text{Median} = 10245$ → 50% des voitures " " 10245 $

$Q_3 = 16515$ → 75% " " (1½)

1.b) Définir et calculer l'étendue.

Etendue = Max(prix) − Min(prix) = 45400 − 5118 = 40282 (½)

1.c) Comment expliquez-vous l'écart entre la moyenne et la médiane ?

(1)

2) Etude du lien entre le prix et la puissance de la voiture

On souhaite étudier le lien entre les deux variables prix et puissance. On suppose que Y=Price et X=puissance. On a les résultats suivants : $s_X = 37.96011$, $s_Y = 8089.083$ et $c_{XY} = 249473.9$.

2.a) Calculer le coefficient de corrélation. Commenter.

$\rho_{XY} = \frac{cov(X,Y)}{S_X \cdot S_Y} = \frac{249473,9}{37,96 \times 8089,083} = 0,81$ alors les deux variables prix et puissance sont fortement positivement liées. (1)

Avec le logiciel R, nous avons ce résultat :

```
>model<-lm(price~horsepower, data=cars)
>Call:
lm(formula = price ~ horsepower, data = cars)

Coefficients:
(Intercept)   horsepower
    -4630.7        173.1

Residual standard error: 4728 on 191 degrees of freedom
Multiple R-squared:  0.6601,Adjusted R-squared:  0.6583
F-statistic: 370.9 on 1 and 191 DF,  p-value: < 2.2e-16
```

2.b) Déterminer l'équation de régression $\hat{Y} = \hat{a}X + \hat{b}$.

$y = 173,1X - 4630,7$ (1)

2.c) Donner et commenter le coefficient de détermination.

$R^2 = 0,6601$: 66% de la variabilité des prix est expliquée par la droite de régression (1)

2.d) Quelle est la prévision pour le prix dont la puissance est 83 ?

X = puissance, Y = prix

prix = 173,1 (83) − 4630,7 = 9736,6 $ (1)

2.e) Si on inverse le rôle de X et Y, est-ce que le coefficient de corrélation est modifié ? Pourquoi ?

Non, car $\rho_{(Y,X)} = \frac{cov(Y,X)}{S_Y \cdot S_X} = \frac{cov(X,Y)}{S_X \cdot S_Y}$ (1)

Exercice 3 (7 ½ pts)

Pour cet exercice, nous allons utiliser le jeu de données Films.txt que vous aviez à travailler en autonomie.

On rappelle que le jeu de données est constitué des 50 succès du box-office français entre 1945 et 2015 caractérisés par 8 variables :

Nom	Echelle
Genre	Comedie, Aventure, Animation, Action, Drame
Nationalite	FR, GB, US, NZ, IT
Entrees	entre 7 380 526 et 21 773 383
Cout	entre 500 000 et 245 000 000
Recette	entre 2 000 000 et 2 731 068 853
Duree	entre 70 et 238
Prix	entre 0 et 208
Annee	entre 1945 et 2015

Une analyse en composantes principales a été effectuée avec les variables quantitatives. Les résultats se trouvent en annexe.
L'axe 1 explique 56.37% de l'inertie du nuage de points et l'axe 2 explique 21.43%.

1) Combien y-a-t-il d'axes en tout ?

5 variables ⇒ 5 axes en total. (½)

2) Quelles variables contribuent le plus à la constitution de l'axe 1 ? de l'axe 2 ?

Axe 1 : coût, recette, prix

Axe 2 : Entree, durée, (on peut dire coût aussi) (2)

3) Sur quel(s) axe(s) peut-on interpréter le film Avatar ? Comment peut-on le qualifier ?

• Sur l'axe 1. (1 ½)

4) Sur quel(s) axe(s) peut-on interpréter le film « Autant en emporte le vent » ? Comment peut-on le qualifier ?

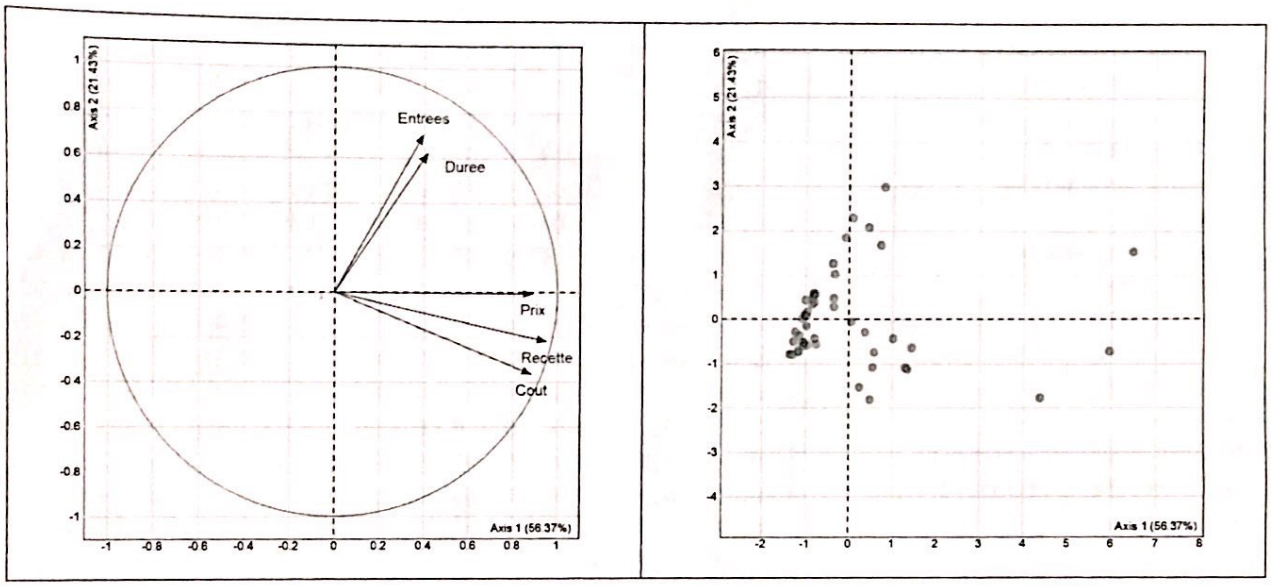
• Sur l'axe 2. (1)

5) Quelle est la contribution moyenne d'un film ? Y-a-t-il des films qui ne respectent pas cette valeur moyenne ? Quel impact cela peut-il avoir ?

- Contribution moyenne d'un film est: $\frac{1}{50} = 0.02$
- Les 5 derniers films ne respectent pas (2)
ces observations pourraient être atypiques et influer sur la construction des axes.

ANNEXE

Résultats de l'ACP



Résultats concernant les variables sur les 3ers axes

$cos2

	Dim.1	Dim.2	Dim.3
Entrees	0,159	0,499	0,330
Cout	0,783	0,135	0,002
Recette	0,909	0,049	0,002
Duree	0,175	0,389	0,433
Prix	0,793	0,000	0,002

$contrib

	Dim.1	Dim.2	Dim.3
Entrees	5,63	46,54	42,96
Cout	27,78	12,58	0,20
Recette	32,26	4,57	0,24
Duree	6,20	36,31	56,34
Prix	28,12	0,00	0,27