	<p style="text-align: center;"><b>ING1-GI</b> EXAMEN DE DATA EXPLORATION 2021-2022</p>
2h	Examen papier 2 feuilles R/V manuscrites autorisées Calculatrice autorisée

## Exercice 1

On a interrogé une partie des élèves d'un collège pour connaître la distance qu'ils doivent parcourir pour se rendre à l'établissement scolaire et qui représente la variable X regroupée selon trois catégories *courte*, *moyenne* et *longue* (i.e. distance domicile/collège).

On s'intéresse de plus à la variable Y qui représente le niveau scolaire de l'élève. L'objectif est d'étudier l'éventuel impact de la distance domicile/collège sur les résultats. On obtient ainsi le tableau suivant :

	Faible	Moyenne	Elevé
Courte	23	25	79
Moyenne	83	85	55
Longue	102	21	27

- 1) Préciser les variables étudiées ainsi que leur type. Quelle est la population étudiée ? Quelle est la taille de l'échantillon ?

Les deux variables sont : la distance parcourue par l'élève pour se rendre à l'établissement scolaire X et son niveau scolaire Y.

Les deux variables sont qualitatives.

La taille de l'échantillon est : 500 élèves.

- 2) Etablir le tableau complet des effectifs observés en ajoutant les effectifs marginaux.

	Faible	Moyenne	Elevé	Total
Courte	23	25	79	127
Moyenne	83	85	55	223
Longue	102	21	27	150
Total	208	131	161	500

- 3) Calculer le tableau des fréquences observées.

	Faible	Moyenne	Elevé	Total
Courte	0.046	0.05	0.158	0.254
Moyenne	0.166	0.17	0.11	0.446
Longue	0.204	0.042	0.054	0.3
Total	0.416	0.262	0.322	1

- 4) Donner le tableau des profils lignes.

	Faible	Moyenne	Elevé	Total
Courte	0.1811	0.196	0.622	1
Moyenne	0.372	0.381	0.246	1
Longue	0.68	0.14	0.18	1

- 5) Comparer le tableau de profils lignes avec le profil moyen ligne. Que pouvez-vous en conclure sur le lien entre les deux variables ? Justifiez votre réponse.

Il faut comparer les lignes de tableau des profils lignes avec le profil moyen ligne suivant :

0.416	0.262	0.322
-------	-------	-------

On peut remarquer que les profils lignes s'écartent du profil moyen ligne. Par exemple toutes distances confondues, on constate qu'il y a 41,6% d'élève ayant un niveau faible. Or pour les distances courtes, ce pourcentage tombe à 18,11%, et a contrario, pour les distances longues, il augmente à 68%. Ce qui nous laisse conclure que les deux variables sont liées. Conclusion à confirmer avec le test de  $\chi^2$ .

- 6) Quelle est la probabilité que :
- un élève parcourt une distance longue ?  
0.3
  - un élève qui parcourt une longue distance ait un niveau faible ?  
0.68 (Tableau des profils lignes)
  - un élève ait un niveau faible et parcourt une distance moyenne ?

0.166 (Tableau des fréquences)

- 7) En supposant que les deux variables sont indépendantes, calculer l'effectif théorique des élèves de niveau faible et qui parcourent une distance courte. Détailler vos calculs.

L'effectif théorique est :  $(208 \times 127) / 500 = 52.832$

- 8) La distance du chi-deux entre le tableau des effectifs observés et celui des effectifs théoriques est 114.74. Peut-on conclure que les variables sont liées ?

Voici le tableau de seuils de décision

d.d.l.	4	5	6	7	8	9	10	15
Seuil	9.49	11.075	12.59	14.07	15.51	16.92	18.31	24.99

Le d.d.l est :  $(3-1) \times (3-1) = 4$  donc le seuil est 9.49..  $\chi^2 = 114.74 \gg 9.49$  alors les variables sont fortement liées.

Attention, les résultats de ce test ne sont valables que si les effectifs théoriques sont  $> 5$  (ce qui doit être le cas car tous les effectifs marginaux sont élevés)

## Exercice 2

Le tableau ci-après présente une partie des offres locatives de 31 appartements rennais proposés sur le site *leboncoin.fr* au 02/01/2017.

### A- Modèle de prévision du loyer en fonction de la surface

En premier temps, on s'intéresse à étudier la relation entre le loyer Y de ces appartements et leur surface X. Avec le logiciel R nous avons obtenu ce résultat :

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  244.296     54.890   4.451 0.000116 ***
Surface       7.282       1.149   6.339 6.31e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 113.7 on 29 degrees of freedom
Multiple R-squared:  0.5808,    Adjusted R-squared:  0.5664
F-statistic: 40.18 on 1 and 29 DF,  p-value: 6.31e-07

```

- 1) Donner une estimation du coefficient de corrélation entre le loyer et la surface d'un appartement. Est-ce que la surface joue un rôle sur le prix de location des appartements ? Considérez-vous ce rôle comme important ?

Le coefficient de détermination est 0.5808, alors le coefficient de corrélation est  $\pm 0.762$ , les variables sont donc fortement corrélées entre elles. Logiquement, on peut dire que le coefficient de corrélation est 0.762 ( et pas - 0.762) car les deux variables Surface et Prix de location sont souvent corrélé de sens positive.

- 2) Proposer un modèle permettant de donner la relation entre le loyer des appartements et leur surface.

X : Surface

Y : Loyer

$$Y = 7.282 X + 244.296 + \epsilon \text{ (ou bien } Y^{\wedge})$$

$$\text{Loyer} = 7.282 * \text{Surface} + 244.296 + \epsilon$$

- 3) Donner le coefficient de détermination. Interpréter.

$R^2 = 0.5808$ , alors 58% de la variabilité de Loyer est expliqué par la droite de régression.

- 4) Un ami qui est en location à Rennes dans un appartement de 30 m<sup>2</sup> m'affirme qu'il paye un loyer de 410 euros. Si je cherche à louer sur Rennes un appartement de même surface :

- a) Calculer la valeur prédite du prix de location.

Pour Surface  $X = 30 \text{ m}^2$ , la valeur de Loyer prédite par le modèle est :  $7.282 * 30 + 244.296 = 462.756$

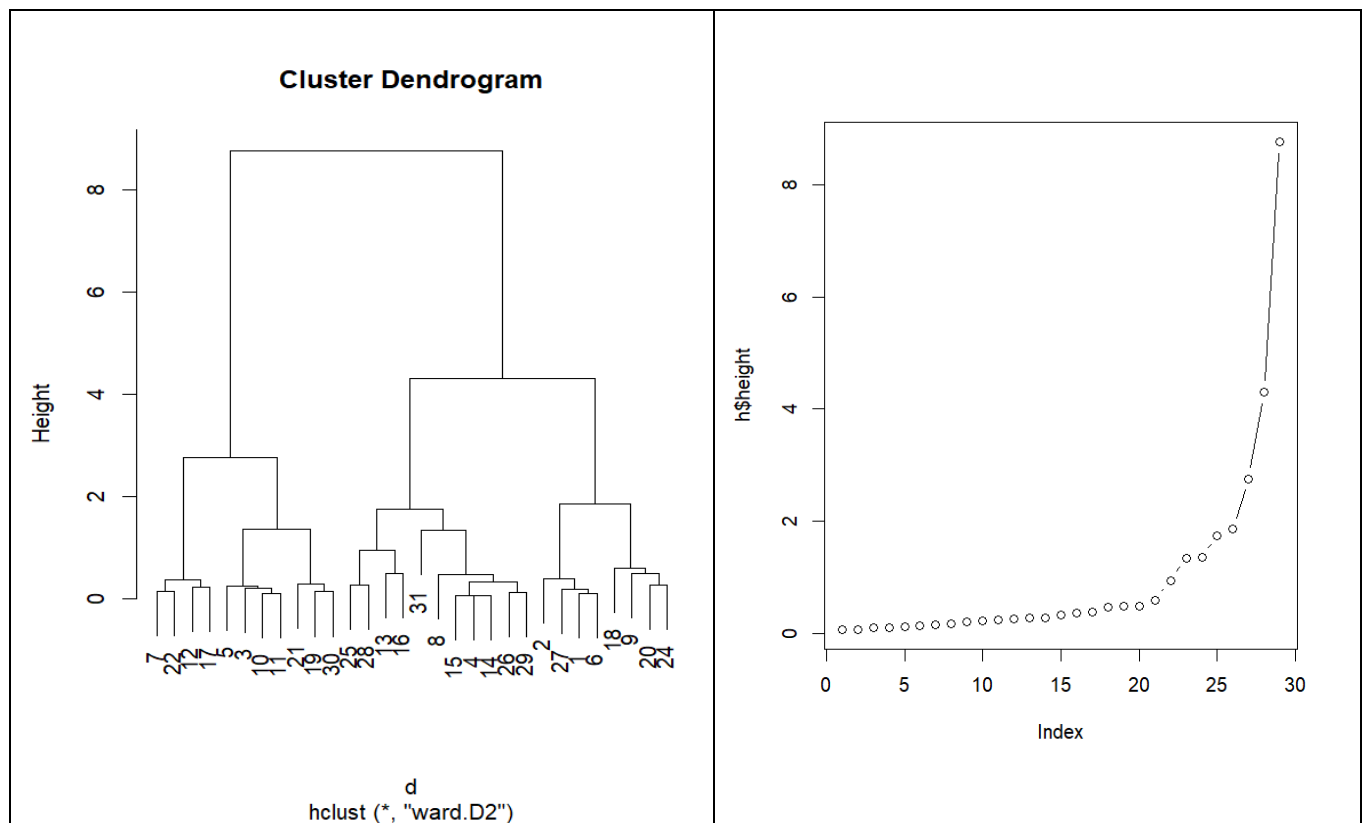
- b) Calculer en détaillant la valeur du résidu dans ce cas.

$$\text{Résidu} = \text{Vrai valeur} - \text{Valeur prédite} = 410 - 462.756 = -52.756$$

RÉFÉRENCE	SURFACE (en m <sup>2</sup> )	TYPE	LOYER (en €)	QUARTIER
1068307364	63	T3	660	Atalante Beaulieu
1072846037	65	T3	610	Le Blosne
1072839939	30	T1	410	Patton
1072833908	47	T2	605	Nord St Martin
1072833680	25	T1	400	Centre
1072825556	65	T3	660	Jeanne d'Arc
1072815333	14	T1	341	Sud-Gare
1072807647	41	T2	570	Chézy-Dinan
1072806540	72	T3	790	La Mabilais
1072850534	28	T1	420	Nord St Martin
1072804498	26	T1	420	St Hélier
1072797911	16	T1	310	Beaulieu
1072796804	54	T3	540	Sud-Gare
1072792504	46	T2	600	Cleunay
1072786440	46	T2	610	Nord St Martin
1062080985	62	T2	510	Bréquigny
1072779442	12	T1	320	La Mabilais
1072766364	63	T4	800	Centre
1072765957	32	T2	507	Ste Thérèse
1064945767	71	T3	737	Landry
1072779613	37	T1	500	Bourg-Levesque
1072779761	15	T1	360	Sud-Gare
1064446623	58	T2	1150	Centre
1072756389	66	T3	740	Longchamps
1072745034	48	T2	500	Centre
1072744005	48	T2	570	Poterie
1047805608	65	T3	640	Francisco-Ferrer
1042429291	44	T2	521	Villejean
1067924582	47	T2	585	Villejean
1072733094	33	T1	490	Centre
1036186345	36	T2	710	Beaulieu

## B- Classification des appartements en fonction de leur loyer et leur surface

On s'intéresse à une classification des appartements à partir de leur loyer et leur surface. Avec le logiciel R, on obtient le dendrogramme suivant :



- 1) De quelle méthode s'agit-il et quelle est la distance utilisée ?  
Classification Ascendante Hiérarchique (CAH), la distance utilisée est celle de Ward.
- 2) Quels découpages peut-on envisager d'après ce dendrogramme ? Expliquer votre choix ?  
4 classes – ou 3 classes pas évident – On peut accepter les deux réponses.
- 3) Proposer une autre méthode de classification et donner la différence entre les deux méthodes

Méthode de K-means. Avec la méthode de K-means on précise dès début le nombre des classes voulu. De plus la méthode K-means est stochastique et on doit l'itérer plusieurs fois avec des initialisations différentes. La méthode CAH est basée sur une distance entre clusters alors que la méthode k-means est basée sur une distance entre points.

- 4) Est-ce que les clusters correspondent aux types de appartements ? Justifier votre réponse. Quelle méthode statistique pouvez-vous utiliser pour confirmer votre réponse ?

Les clusters ne correspondent pas aux types de appartements, car dans chaque classe on peut trouver des appartements de différents Types.

On peut utiliser un test du chi-deux entre la variable types d'appartement et le variable classes.