



# Examen Data Exploration - ING1 GM-MIM

2 feuilles R/V manuscrites autorisées, Calculatrice autorisée

Date : 22 décembre 2023

Durée : 2h

Nombre de pages : 5

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

◇ ◇ ◇

## Exercice 1

Le jeu des données Airbnb se compose d'informations publiques sur les annonces et les mesures Airbnb à New York. Les données ouvertes Airbnb de la ville de New York 2019 comprennent des informations sur environ 28590 logements Airbnb dans la ville et sont mises à la disposition du public par le gouvernement de la ville de New York pour promouvoir la transparence et la compréhension de l'impact des locations sur la ville. Dans la suite, on prendra que les 500 premiers logements et voici un extrait :

Id	Neighbourhood	Latitude	Longitude	Room type	Price\$	Minimum night	Number of reviews	Availability
2539	Brooklyn	40.64747	-73.9724	Private room	125	5	149	324
2595	Manhattan	40.15792	-73.5896	Entire home	225	3	45	194
3647	Brooklyn	40.76489	-73.965	Shared room	53	15	21	49
...	...	...	...	...	...	...	...	...

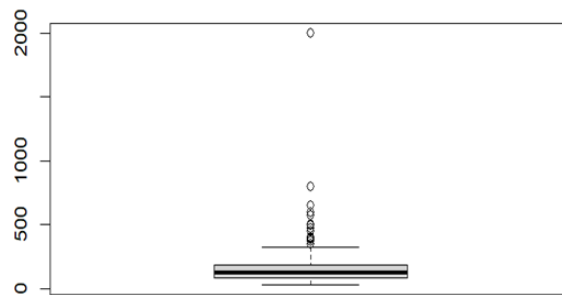
1. Préciser les variables étudiées, donner leur type et un moyen de les représenter graphiquement
2. Quelle est la population étudiée ? Quelle est la taille de l'échantillon ?

## Analyse Univariée : Étude de la variable prix (en dollars)

Le logiciel R nous fournit les résultats suivants :

```
>quantile (Airbnb$price)
0%   25%   50%   75%  100%
33   85   125   186  2000
```

3. Donner et interpréter les quartiles.
4. Calculer l'écart inter-quartiles et l'étendue.
5. La moyenne de prix par nuit est de 152.3\$. Comment interprétez-vous la différence entre la moyenne et la médiane ?
6. Parmi les indicateurs numériques calculés avant, lesquels sont des indicateurs de dispersion et lesquels sont des indicateurs de positions ?
7. Voici le Boxplot (boîte à moustache) pour la variable 'prix'.



- (a) Calculer les extrémités des moustaches.
- (b) Y a-t-il des valeurs aberrantes ? Justifier votre réponse.

### Analyse Bivariée : Etude entre Prix et Neighbourhood (arrondissement)

La ville de New York compte 5 arrondissements (neighbourhood). On s'intéresse maintenant à étudier le lien entre les deux variables arrondissements et le prix des locations dans ces arrondissements. Le logiciel R nous fournit les résultats suivants :

```
> model <- lm (price~as.factor(neighbourhood),data=Airbnb)
> anova(model)
```

Analysis of Variance Table

Response: price

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
as.factor(neighbourhood)	68	1253070	18428	1.1695	0.1819
Residuals	431	6791083	15757		

8. Donner la variance inter groupes et la variance intra groupes. En déduire la variation totale.
9. Calculer le rapport de corrélation et interpréter cette valeur.

### Analyse Bivariée : Etude entre Room-Type (type de chambre) et Neighbourhood (arrondissement)

On donne le tableau de contingence

	Entire home/apt	Private room	Shared room
Bronx	3	4	0
Brooklyn	139	98	0
Manhattan	139	86	4
Queens	10	12	0
Staten Island	0	5	0

10. Etablir le tableau complet des effectifs observés en ajoutant les effectifs marginaux.
11. Calculer et interpréter les valeurs suivantes  $n_{.1}$ ,  $f_{2|2}$ .
12. Donner le tableau de profils colonnes. Comparer le tableau de profils colonnes avec le profil moyen colonne. Que pouvez-vous prédire sur le lien entre ces deux variables ?

On donne le tableau des effectifs théoriques

	Entire home/apt	Private room	Shared room
Bronx	4.074	2.87	0.056
Brooklyn	137.934	97.17	?
Manhattan	133.278	93.89	1.832
Queens	12.804	9.02	0.176
Staten Island	2.910	?	0.04

13. Compléter les valeurs manquantes dans le tableau des effectifs théoriques, en détaillant les calculs.
14. Le calcul de la distance du khi-deux nous donne la valeur 15.139. Expliquer comment cette distance est calculée.
15. Y a-t-il un lien entre l'arrondissement et le type de chambre ? Justifiez votre réponse.

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

16. Faites vous confiance à ce résultat ? Justifiez votre réponse.

## AFC

Maintenant, on effectue le calcul sur tout le data (28590 logements), et on trouve résultat suivant

```
> res = chisq.test(table(airbnb$neighbourhood_group,airbnb$room_type))
```

Pearson's Chi-squared test

data: A

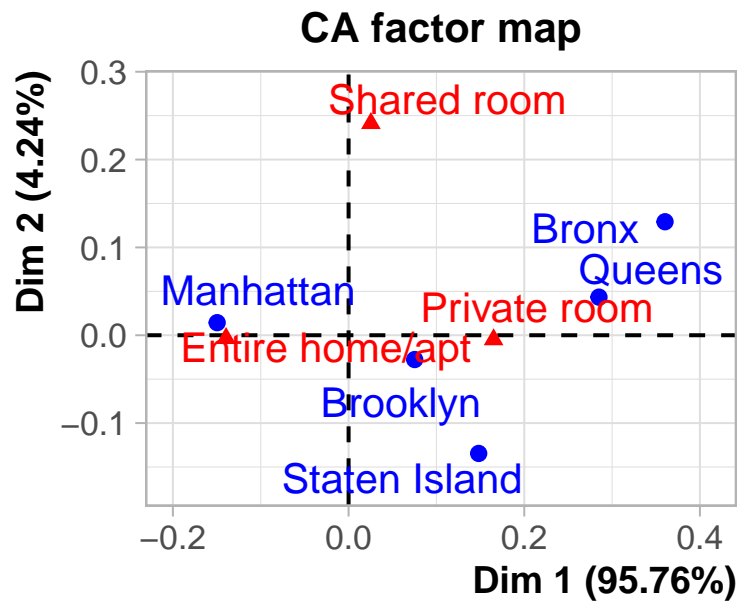
X-squared = 675.85, df = 8, p-value < 2.2e-16

17. Est-ce que la réponse de question 15 va changer ? Justifier votre réponse.

On effectue un AFC

```
> res.ca$eig
      eigenvalue      percentage of variance  cumulative percentage of variance
dim 1 0.022636919          95.758815          95.75881
dim 2 0.001002596           4.241185          100.00000
```

18. Que représente le pourcentage affiché sur les axes ? Expliquez pourquoi la somme des trois axes fait exactement 100%.
19. Interprétez le graphique.



## Exercice 2

On considère data : *Climfrance.txt*. Ces données se composent de 36 lignes (observations) et 11 colonnes (caractéristiques/variables). Les caractéristiques sont : l'altitude, la latitude, la longitude, la température annuelle moyenne, la température annuelle maximale, la température annuelle minimale, l'humidité relative, les précipitations annuelles moyennes, les précipitations maximales en 24 heures, le nombre de jours de pluie et le nombre d'heures d'ensoleillement par an.

```
clim = read.table("Climfrance.txt", header = T, sep = ";", dec = ".")
install.packages("FactoMineR")
library(FactoMineR)
res.PCA.clim <- PCA(clim)
head(res.PCA.clim$eig)
```

##	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.6985564	42.714149	42.71415
comp 2	2.9427020	26.751837	69.46599
comp 3	1.3401647	12.183316	81.64930
comp 4	0.7815300	7.104818	88.75412
comp 5	0.5514082	5.012802	93.76692
comp 6	0.2385331	2.168483	95.93540

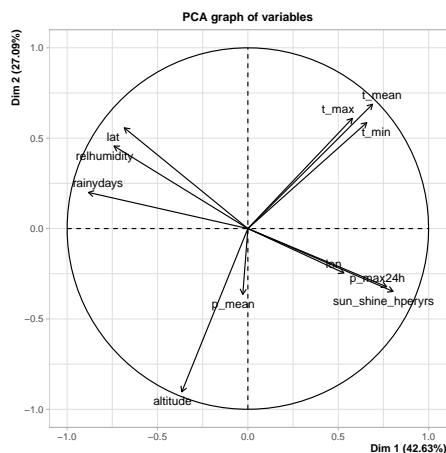
```
> res.PCA.clim$var$contrib[,1:4]
```

	Dim.1	Dim.2	Dim.3	Dim.4
altitude	2.84792993	27.380371	0.1788591	1.15421895
lat	9.97555726	10.421259	4.7106116	1.84063577
lon	6.02050855	2.025628	17.3696828	36.95051294
t_mean	10.14368290	15.869571	1.8020756	0.07607592
t_max	7.13248282	12.461242	0.1501642	14.19652610
t_min	9.22985599	11.526418	2.4517041	8.83405789
relhumidity	11.69989549	6.993222	3.0985483	3.20138110
p_mean	0.01680045	4.448227	56.7672194	9.42186140
p_max24h	12.57535176	3.476664	8.3729141	2.68703808
rainydays	16.59068209	1.338614	4.8983872	4.76582364
sun_shine_hperyrs	13.76725277	4.058785	0.1998335	16.87186820

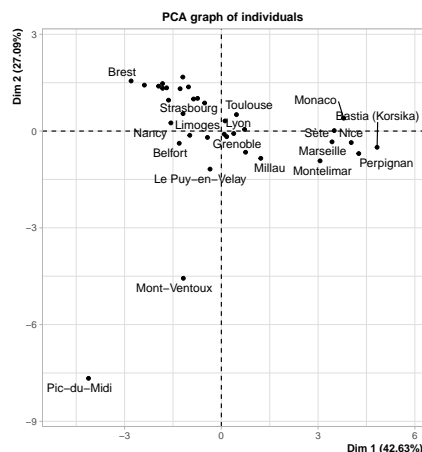
```
> res.PCA.clim$var$cos2[,1:4]

          Dim.1      Dim.2      Dim.3      Dim.4
altitude  0.1335494714  0.81604125  0.002265643  0.009432545
lat        0.4677890372  0.31059394  0.059670231  0.015042103
lon        0.2823228641  0.06037157  0.220025144  0.301968164
t_mean     0.4756730411  0.47297476  0.022827241  0.000621710
t_max      0.3344672570  0.37139335  0.001902159  0.116017305
t_min      0.4328204769  0.34353196  0.031056211  0.072193971
relhumidity 0.5486493340  0.20842513  0.039249913  0.026162429
p_mean     0.0007878324  0.13257442  0.719081386  0.076997637
p_max24h    0.5897025642  0.10361807  0.106061328  0.021959099
rainydays  0.7779955548  0.03989588  0.062048822  0.038947416
sun_shine_hperyrs 0.6455950031  0.12096755  0.002531330  0.137880821

plot(res.PCA.clim, choix="var")
```



```
plot(res.PCA.clim, choix="ind")
```



En regardant le résultat de la fonction PCA, répondre aux questions suivantes :

1. Combien y-a-t-il d'axes au total ? Justifiez votre réponse.
2. Combien d'axes peut-on choisir pour interpréter le résultat ? Pourquoi ? A quoi correspond la somme des valeurs propres ?
3. Quelle est la contribution moyenne d'une variable contribuant à la construction des axes ? Quelles variables contribuent majoritairement à la deuxième axe ?
4. Interpréter les axes, la liaison entre les variables.
5. Y a-t-il des variables qui sont mal représentées sur le plan principale ? Justifiez vos réponse. Si oui, elles sont bien représentées sur quel(s) axe(s) ?
6. Interpréter les caractéristiques de ville "Brest" et "Lyon" et "Pic-du-Midi".