Analyse des Données - TP2

Quentin Garnier

2025

1 Criminalité aux USA

1.1 Exploration élémentaire

1.1.1 Identification des valeurs atypiques

Pour détecter des valeurs atypiques dans notre jeu de données, nous analysons d'abord les statistiques descriptives. En observant la variable **robbery**, nous constatons que :

- Le **3ème quartile (Q3)** est de 155.85.
- La valeur **maximale** est de 472.0.

Afin de valider cette observation, nous afficons le **boxplot** de la variable *robbery*. Comme prévu, nous voyons un point isolé au-dessus de la moustache supérieure, ce qui confirme la présence d'une valeur extrême.

Cette valeur pourrait correspondre à un état où les taux de vols avec violence sont exceptionnellement élevés. Une anlyse plus approfondie permettrait d'identifier les causes sous-jacentes de cette anomalie.

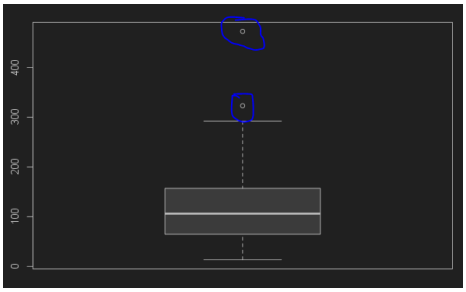

Figure 1: BoxPlot des données de robbery

L'analyse de la matrice de corrélation nous permet d'identifier des relations entre les différentes variables criminelles.

1.1.2 Corrélations fortes

Certaines variables montrent une corrélation élevée, indiquant une relation sgnificative :

- **Rape et Assault** : 0.74 (forte corrélaton positive)
- **Burglary et Larceny** : 0.79 (forte corrélation positive)
- **Rape et Burglary** : 0.71 (forte corrélation positivbe)

1.1.3 Corrélations faibles ou inexistantes

Certaines variables ne montrent pas de lien significatif :

- **Murder et uto** : 0.06

1.1.4 Interprétation

L'analyse révèle deux grands groupes :

- **Les crimes violents** (Murder, Rape, Assault, Robbery) sont fortement corrélés entre eux.
- **Les crimes contre les biens** (Burglary, Larceny, Auto) montrent également une forte corrélation.
- Il y a peu de lien entre ces deux goupes

1.2 Analyse en Composantes Principales (ACP)

1.2.1 Réalisation de l'ACP

Nous effectuons l'ACP sur la table `crime`

1.2.2 Choix de la dimension

L'analyse des corrélations montre deux groupes principaux :

- **Crimes violents** : Murder, Rape, Assault, Robbery
- **Crimes contre les biens** : Burglary, Larceny, Auto

1.2.3 Représentation des individus et vriables

1.2.4 Variance des composantes

Le graphique ci-dessous montre la variance expliquée par chaque composante :

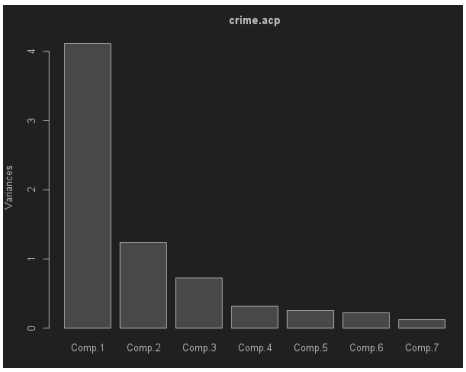

Figure 2: BoxPlot des données de robbery

Observation : Les deux premières composantes expliquent l'essentielde la variance, les suivantes apportent peu d'information.

1.2.5 Confirmation du choix de la dimension

Les biplots permettent d'évaluer la répartition des individus et des variables :

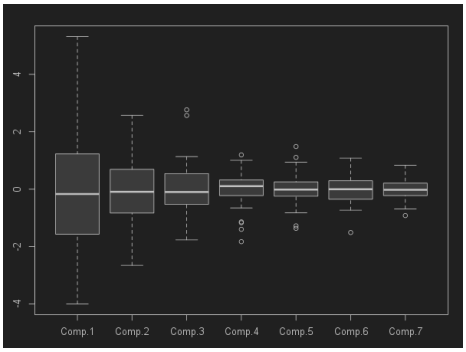

Figure 3: BoxPlot des données de robbery

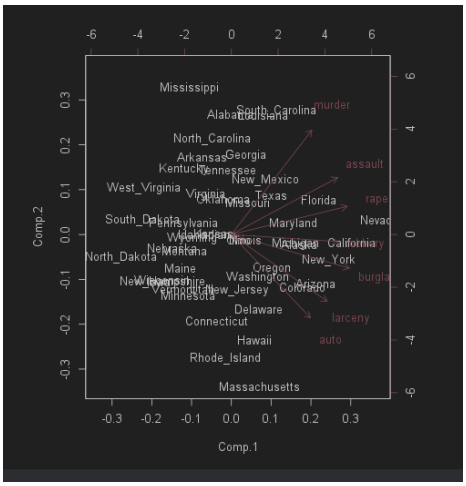

Figure 4: BoxPlot des données de robbery

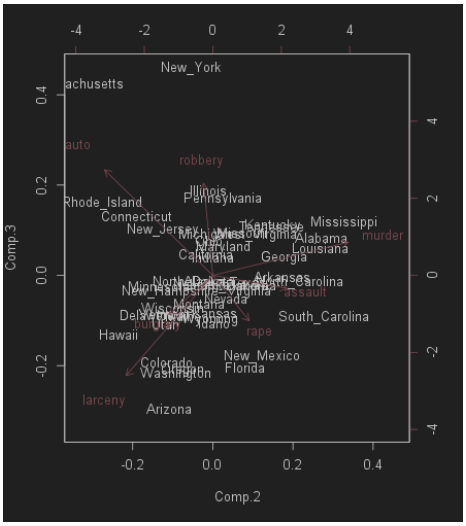

Figure 5: BoxPlot des données de robbery

Conclusion : Les composantes 1et 2 suffisent pour bien représenter les données. La troisième n'apporte pas d'information significative supplémentaire.

1.2.6 Identification des valeurs atypiques sur la troisième composante

D'après l'ACP, deux états apparaissent comme **atypiques** sur la troisième composante : **New York** et **Massachusetts**.

Ces valeurs influencent potentiellement la définition des axes. Il est recommandé de vérifier si leur exclusion modifie l'interprétation des résultats.

1.2.7 Exclusion des valeurs atypiques

Après avoir exclu **New York** et **Massachusetts**, vici les nouveaux biplots :

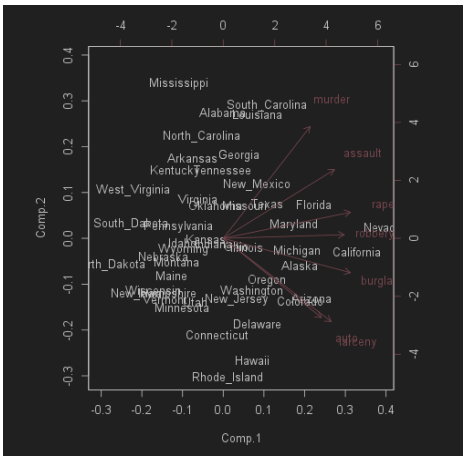

Figure 6: Biplot des composantes 1 et 2 après exclusion

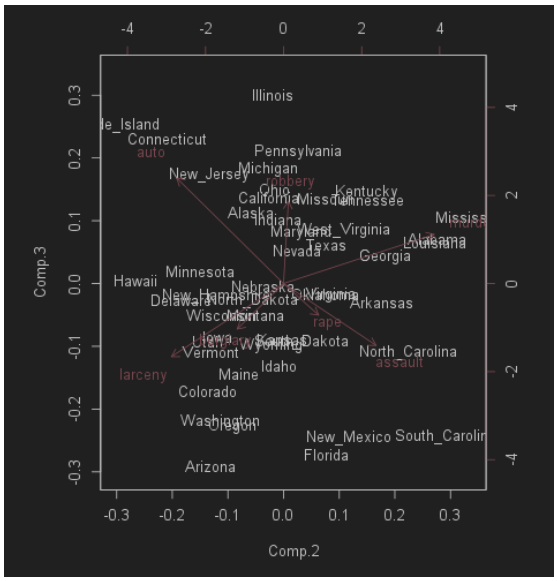

Figure 7: Biplot des composantes 2 et 3 après exclusion

Observations :

- Le plan (**Comp.1, Comp.2**) reste similaire après l'exclusion.
- L'axe **Comp.3** est légèrement modifié,ce qui montre que ces états avaient une influence sur cette dimension.

Conclusion : L'exclusion n'affecte pas l'analyse globale mais réduit l'effet des valeurs extrêmes.

1.2.8 Interprétation des axes

je ne comprends pas cette question

2 Hôtels Méditerranéens

2.1 ACP avec une variable qualitative supplémentaire

2.1.1 Chargement des données

2.1.2 Analyse des variables

Le jeu de données `hotels` contient 39 hôtels et 8 variables :

- **PAYS** : Variable qualitative indiquant le pays dimplantation (variable supplémentaire dans l'ACP).
- **ETOILE** : Nombre d'étoiles de l'hôtel (quantitative discrète).
- **CONFORT** : Niveau de confort noté sur une échelle (quantitative discrète).
- **CHAMBRE** : Nombre total de chambres (quantitative continue).
- **CUISINE** : Niveau d'équipement de la cuisine (quantitative discrète).
- **SPORT** : Niveau d'équipement sortif (quantitative discrète).
- **PLAGE** : Indicateur d'accès à une plage (quantitative discrète).
- **PRIX** : Prix moyen des chambres (quantitative continue).

Individus analysés : Chaque ligne du tableau correspond à un hôtel.
Statistiques descriptives :

PAYS	ETOILE	CONFORT	CHAMBRE	CUISINE	SPORT
Length:39	Min. :0.000	Min. :2.000	Min. : 50.0	Min. : 1.000	Min. : 0.000
Class :character	1st Qu.:2.000	1st Qu.:4.000	1st Qu.:148.0	1st Qu.: 5.000	1st Qu.: 4.000
Mode :character	Median :3.000	Median :5.000	Median :250.0	Median : 7.000	Median : 6.000
	Mean :2.974	Mean :5.179	Mean :261.2	Mean : 6.667	Mean : 6.231
	3rd Qu.:4.000	3rd Qu.:6.000	3rd Qu.:317.0	3rd Qu.: 9.000	3rd Qu.:10.000
	Max. :5.000	Max. :9.000	Max. :800.0	Max. :10.000	Max. :10.000

PLAGE	PRIX
Min. : 0.000	Min. : 369.0
1st Qu.: 6.500	1st Qu.: 447.0
Median : 8.000	Median : 495.0
Mean : 7.769	Mean : 529.9
3rd Qu.:10.000	3rd Qu.: 574.0
Max. :10.000	Max. :1101.0

Figure 8: Biplot des composantes 2 et 3 après exclusion

Covariances et corrélations :

	ETOILE	CONFORT	CHAMBRE	CUISINE	SPORT	PLAGE	PRIX
ETOILE	1.00000000	0.62945260	0.08047836	0.5956659	0.08291707	-0.12446837	0.54146171
CONFORT	0.62945260	1.00000000	0.07230083	0.5586882	0.03591748	-0.05166764	0.47368666
CHAMBRE	0.08047836	0.07230083	1.00000000	0.4209873	0.47835884	0.18350863	-0.03491169
CUISINE	0.59566586	0.55868815	0.42098735	1.0000000	0.45852342	0.25575255	0.56748515
SPORT	0.08291707	0.03591748	0.47835884	0.4585234	1.00000000	0.53124345	0.31135324
PLAGE	-0.12446837	-0.05166764	0.18350863	0.2557525	0.53124345	1.00000000	0.33712188
PRIX	0.54146171	0.47368666	-0.03491169	0.5674851	0.31135324	0.33712188	1.00000000

Figure 9: Biplot des composantes 2 et 3 après exclusion

On observe :

- Une **forte corrélation linéaire** entre **ÉTOILE et CONFORT** ($r = 0.63$).
- Une **corrélation notable** entre **ÉTOILE et CUISINE** ($r = 0.60$).
- Une **corrlation moyenne** entre **ÉTOILE et PRI** ($r = 0.54$), ce qui est attendu puisque les hôtels plus étoilés sont souvent plus chers.
- Peu de corrélations fortes entre les autres variables.

Identification des liaisons non linéaires : Si certaines relations ne sont pas linéaires, elles peuvent être visualisées avec :

```
pairs(hotels[,sapply(hotels, is.numeric)])
```

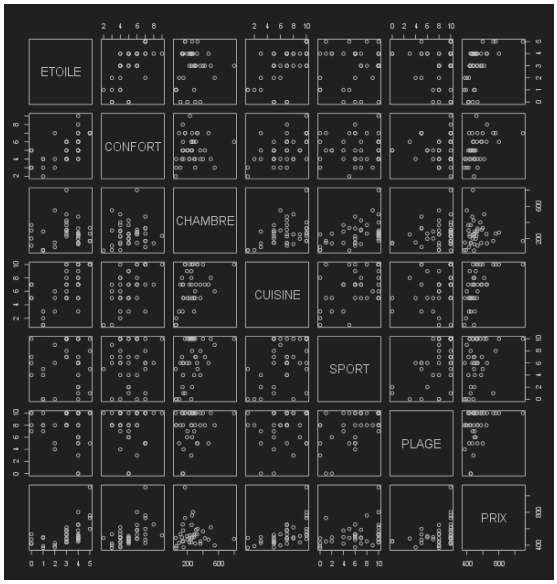

Figure 10: Biplot des composantes 2 et 3 après exclusion

On observe des structures non linéaires, notamment entre :

- **PRIX et CUISINE**
- **SPORT et d'autres vaiables**
- **PLAGE et PRIX**

2.1.3 Analyse en Composantes Principales

Les diagrammes des valeurs propres montrent l'inertie expliquée par chaque composante :

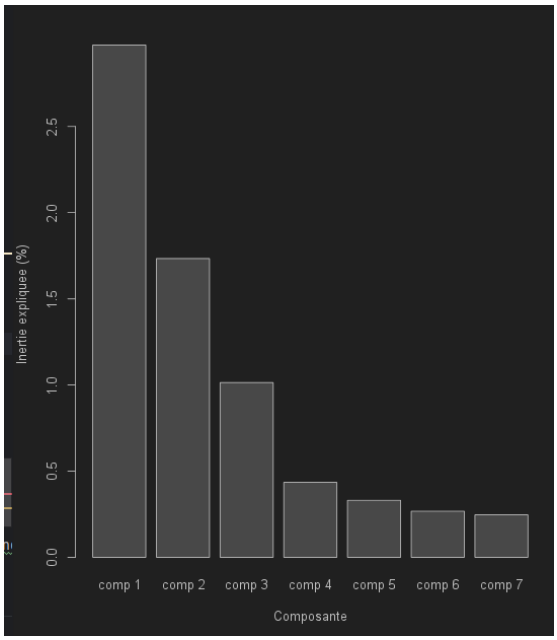

Figure 11: Inertie expliquée par composante

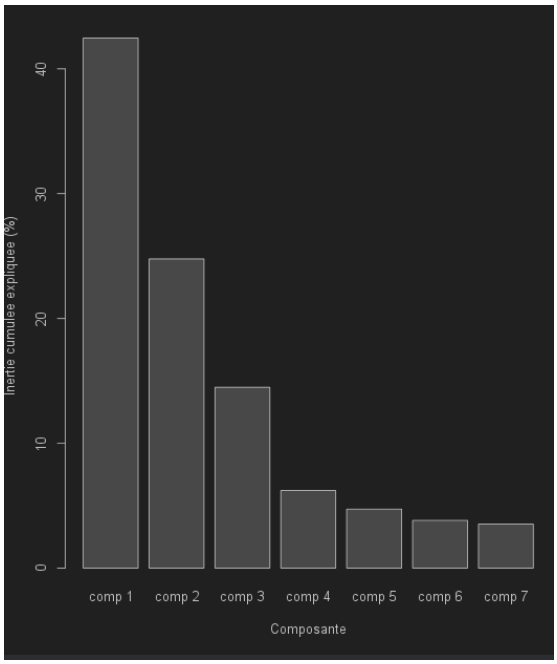

Figure 12: Inertie cumulée expliquée

L'inertie cumulée dépasse **40% après la troisième cmposante** mais reste loin du seuil de **80%**. Le choix optimal dépend du compromis entre simplification et information retenue, mais **3 composantes principales semblent suffisantes pour un bon résumé des données**.

2.1.4 Représentation des individus et des variables

Les individus (hôtels) sont représentés sur les premiers axes factoriels :

Figure 13: Projection des individus sur Dim 1 et Dim 2

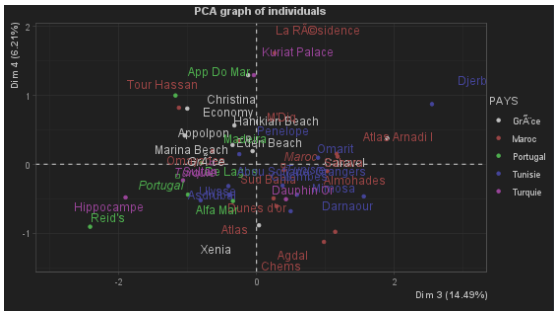

Figure 14: Projection des individus sur Dim 3 et Dim 4

Les variables sont représentées dans les cercles de corrélation :

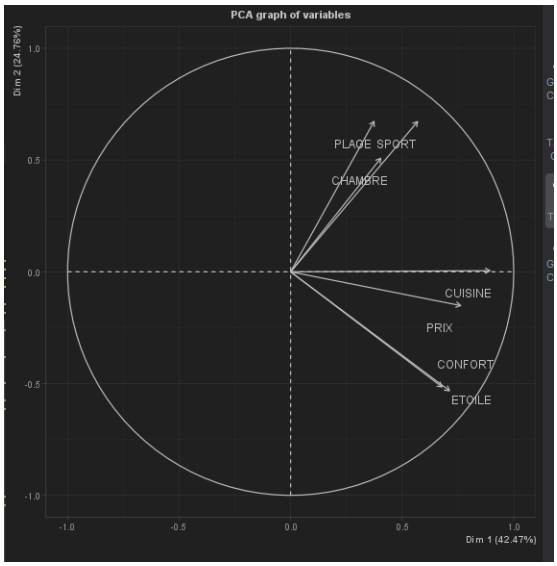

Figure 15: Cercle des corrélations - Dim 1 et Dim 2

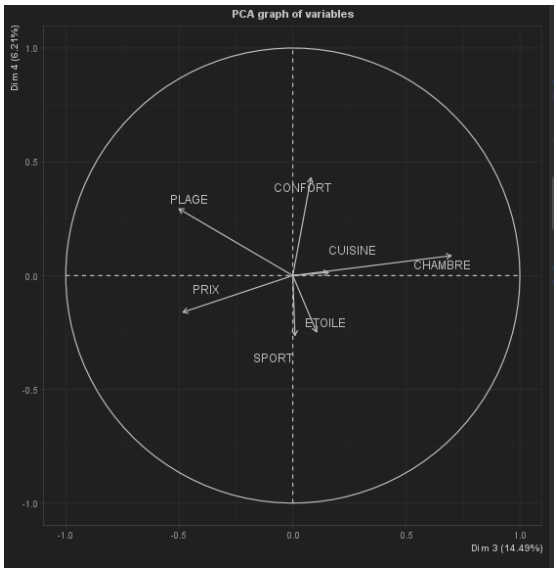

Figure 16: Cercle des corrélations - Dim 3 et Dim 4

L'analyse des représentations montre une bonne séparation des hôtels par pays sur les premiers axes et une forte structuration des variables dans les cercles de corrélation.

2.1.5 Utilisation de dimdesc

- **Dimension 1** :
  - Les variables continues les plus corréles sont **CUISINE** (0.89), **PRIX** (0.76) et **ETOILE** (0.71).
  - La variable catégorielle **PAYS** est significative (p-*value* = 0.006).
  - Les pays **Portugal** et **Grèce** inluencent cette dimension.
- **Dimension 2** :
  - Les variables continues les plus corrélées sont **PLAGE** (0.67) et **SPORT** (0.67).
  - La variable **PAYS** est aussi significative (p-*value* = 0.009).
  - Les pays **Tunisie** et **Maroc** influencent cette dimension.
- **Dimension 3** :
  - Les variables coninues les plus corrélées sont **CHAMBRE** (0.69) et **PRIX** (-0.48).
  - La variable **PAYSl** est également significative (p-*value* = 0.011).
  - Le pays **Portugal** influence cette dimension.

2.1.6 Utilisation de coord.ellipse

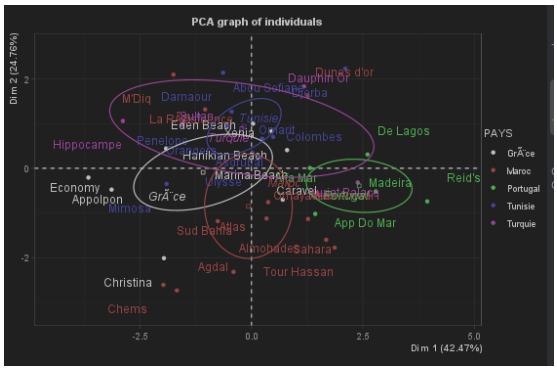

Figure 17: Cercle des corrélations - Dim 3 et Dim 4

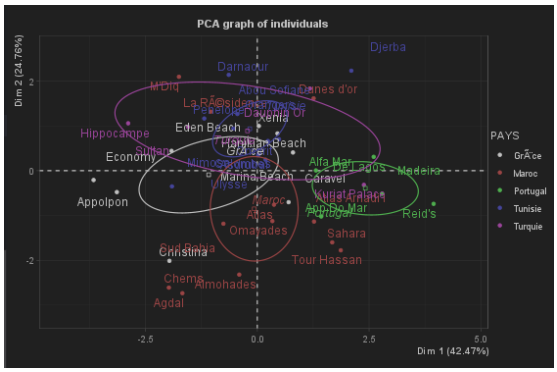

Figure 18: Cercle des corrélations - Dim 3 et Dim 4

On observe que :

- Les hôtels situés au **Portugal** et au **Maroc** sont plutot bien distincts des autres groupes.
- La **Tnisie** et la **Turquie** présentent une forte superposition, suggérant que leurs hôtels partagent des caractéristiques similaires.

2.2 ACP avec variables qualitative et quantitative supplémentaires

Ayant un projet a rendre pour lundi 10 février pour mon universitée eramsus en Italie, je n'ai pu commencer a travailler sur ce projet qu'aujourd'hui, je n'ai donc pas eu le temps de faire cette partie du TP.