PROJET

Analyse Exploratoire de Données

Master 2 Informatique

Université Sorbonne Paris Nord

Louenas Bounia
louenas.bounia@univ-paris13.fr

Année Universitaire 2025-2026

Soutenances : Dernière séance du TP

Datasets

Dataset 1 : Boston Housing (Régression)

Source : Kaggle - Boston Housing Dataset

Variables : 13 features, 506 observations
Cible : Prix médian des maisons (MEDV)

Dataset 2 : Mall Customers (Clustering)

Source : Kaggle - Mall Customers Dataset

Variables : CustomerID, Gender, Age, Annual Income (k$), Spending Score (1-100)
Observations : 200 clients

Travail à réaliser

PARTIE 1 : Boston Housing - Régression

Statistiques descriptives

1. Calculez moyenne, médiane, écart-type, quartiles pour chaque variable
2. Calculez les intervalles de confiance à 95% des moyennes
3. Testez la normalité avec le test de Shapiro-Wilk (`https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html`)
4. Visualisez : histogrammes, boxplots, courbes de densité (KDE)

Analyse bivariée

1. Créez une matrice de corrélation (heatmap)
2. Pour chaque paire : calculez le coefficient r, R^2, et l'IC_{95} du R^2
3. Pour les corrélations significatives ($\alpha = 0.03$) : tracez la droite de régression
4. Identifiez les 3 variables les plus corrélées avec MEDV

Matrice de variance-covariance

1. Calculez la matrice de variance-covariance pour toutes les variables
2. Interprétez les valeurs de variance (diagonale) et de covariance
3. Identifiez les paires de variables avec les covariances les plus élevées
4. Comparez la matrice de covariance avec la matrice de corrélation : quelles différences observez-vous ?
5. Calculez les valeurs propres et vecteurs propres de cette matrice

Analyse en Composantes Principales (ACP)

1. Standardisez les variables et justifiez cette étape
2. Appliquez l'ACP sur toutes les variables (sauf MEDV)
3. Calculez et visualisez la variance expliquée par chaque composante

4. Créez un scree plot et déterminez le nombre optimal de composantes à retenir (critère de Kaiser, coude, variance cumulée $> 80\%$)
5. Visualisez le cercle des corrélations : quelles variables contribuent le plus aux PC1 et PC2 ?
6. Interprétez les deux premières composantes principales : que représentent-elles ?
7. Projetez les observations dans l'espace des 2 premières composantes (biplot)
8. Identifiez d'éventuels outliers dans cet espace réduit

Régression linéaire multiple

1. Sélectionnez 3 à 5 variables pour prédire MEDV et justifiez votre choix
2. Appliquez la régression linéaire multiple
3. Calculez R^2, R^2 ajusté, RMSE
4. Testez la significativité des coefficients (p-values)
5. Visualisez : valeurs prédites vs valeurs réelles
6. Interprétez chaque coefficient : quel est l'impact sur MEDV ?
7. Y a-t-il des variables à retirer ? Justifiez

Régression sur composantes principales (PCR)

1. Appliquez une régression linéaire en utilisant les composantes principales comme prédicteurs
2. Comparez les performances (RMSE, R^2) avec la régression sur variables originales
3. Discutez des avantages et inconvénients de cette approche

PARTIE 2 : Mall Customers - Clustering

Exploration

1. Statistiques descriptives (Age, Income, Spending Score)
2. Matrice de corrélation
3. Normalisez les données

Matrice de variance-covariance et ACP

1. Calculez la matrice de variance-covariance des variables numériques
2. Interprétez les variances et covariances
3. Appliquez l'ACP sur les variables normalisées
4. Déterminez le nombre de composantes à retenir (variance expliquée)
5. Visualisez le cercle des corrélations
6. Projetez les clients dans l'espace des 2 premières composantes
7. Cette projection permet-elle d'identifier visuellement des groupes ?

K-Means

1. Déterminez k optimal avec : méthode du coude + silhouette
2. Appliquez K-Means avec k optimal sur les données originales
3. Appliquez K-Means sur les composantes principales retenues
4. Visualisez en scatter plot : Income vs Spending Score (coloré par cluster)
5. Caractérisez chaque cluster : moyennes, répartition par genre, taille, profil type
6. Comparez les résultats K-Means sur données originales vs sur composantes principales

Classification Hiérarchique Ascendante (CAH)

1. Construisez des dendrogrammes avec : complete, average, Ward
2. Choisissez le meilleur critère et coupez le dendrogramme
3. Matrice de confusion : comparez K-Means vs CAH
4. Les deux méthodes identifient-elles les mêmes groupes ?
5. Appliquez la CAH sur les composantes principales : observe-t-on une meilleure séparation ?

Évaluation et validation

1. Calculez le coefficient de silhouette pour chaque méthode
2. Calculez l'indice de Davies-Bouldin
3. Quelle méthode produit les meilleurs clusters ?
4. L'utilisation de l'ACP améliore-t-elle la qualité du clustering ?

Recommandations business

Pour chaque cluster identifié :

— Profil client

— Produits à proposer

— Stratégie marketing adaptée

Livrables

Groupes de 3 à 4 étudiants

Rapport (PDF, 25 pages maximum)

— Partie 1 : Boston Housing (régression, ACP, variance-covariance)

— Partie 2 : Mall Customers (clustering, ACP, validation)

— Synthèse générale et comparaisons

Code Python

— Scripts commentés et reproductibles

— Jupyter Notebook ou fichiers .py

Présentation (25 slides maximum)

— Régression et ACP : 10 slides maximum

— Clustering et réduction de dimension : 10 slides maximum

— Synthèse : 5 slides

Soumission via Moodle

Fichiers : `NOM1_NOM2_Code.zip`

Soutenance

Durée : 15-20 minutes maximum de présentation + 5-10 minutes de questions

Note finale sur 20 = (Rapport + code $\times$ 0,5) + (Soutenance $\times$ 0,5)