Processus d’analyse exploratoire des données

Titre du cours : Analyse exploratoire des données

Code officiel : 420-A55-BB

Professeur : Dr Komi SODOKE

Plan

I- **Introduction**

II- **Définition de l'analyse exploratoire des données**

III- **Analyse des besoins et sélection des données appropriées**

A. Population, échantillon, estimations, etc.

IV- **Exploration des caractéristiques des données**

A. Unidimensionnelle, multidimensionnelle, séries temporelles, probabilités, etc.

V- **Analyse avancée des données**

A. Composantes Principales, Correspondances simple et multiple, etc.

VI- **Détection des valeurs atypiques**

A. Aberrantes, manquantes, etc.

Plan

VII- Interprétation des résultats

VIII- Documentation

IX- Conclusion

I- Introduction

L'analyse exploratoire des données (AED) est cruciale pour comprendre les données avant des analyses approfondies. Ce chapitre couvre la définition des caractéristiques des données (quantitatives, qualitatives, structurées), l'analyse des besoins et la sélection des données appropriées, l'exploration des caractéristiques des données sous différentes dimensions, l'analyse avancée (ACP, correspondances), la détection des valeurs atypiques, l'interprétation des résultats et la documentation du processus. L'AED combine des méthodes visuelles et statistiques pour fournir une compréhension approfondie, préparer des analyses rigoureuses et des modélisations précises.

II- Définition de la notion de la caractéristique de la donnée

Une donnée peut être définie par différentes caractéristiques qui influencent son analyse et son interprétation.

- ➔ **Nature des données** : Quantitative (numérique) vs Qualitative (catégorique)
- ➔ **Format des données** :
  - ◆ **Données structurées** : Ces données sont organisées dans un format prédéfini, comme les bases de données, les feuilles de calcul. Elles sont faciles à trier et à analyser.
  - ◆ **Données non structurées** : Elles n'ont pas de format prédéfini et comprennent du texte libre, des images, des vidéos. Leur analyse nécessite souvent des techniques avancées comme le traitement du langage naturel ou la reconnaissance d'image.

II- Définition de la notion de la caractéristique de la donnée

➔ **Propriétés statistiques** :

- **Moyenne** : C'est la valeur centrale des données, obtenue en divisant la somme des valeurs par le nombre de valeurs.
- **Médiane** : La valeur médiane sépare les données en deux moitiés égales, la moitié supérieure et la moitié inférieure.
- **Variance** : Mesure de la dispersion des données par rapport à la moyenne.
- **Écart-type** : Racine carrée de la variance, elle indique la dispersion des valeurs par rapport à la moyenne.

Exemple : Une base de données de ventes contient plusieurs types de données :

- **Prix (quantitatif)** : Les montants des ventes exprimés en chiffres.
- **Noms de clients (qualitatif)** : Descripteurs des clients, comme leurs noms, qui ne peuvent pas être mesurés numériquement.
- **Dates d'achat (temporel)** : Information sur le moment où les transactions ont eu lieu.

III- Analyse des besoins et sélection des données appropriées

Avant d'exploiter les données, il est essentiel d'identifier les besoins et de sélectionner les données pertinentes.

➔ **Population et Échantillon**

Dans de nombreux cas, il n'est pas pratique ou possible de travailler sur l'ensemble de la population en raison de contraintes de temps et de ressources. Ainsi, un échantillon représentatif est souvent utilisé. Un échantillon représentatif doit refléter fidèlement les caractéristiques de la population afin que les conclusions puissent être généralisées.

- **Population :** L'ensemble total des sujets ou éléments d'intérêt dans une étude.
- **Échantillon :** Une sous-ensemble de la population sélectionnée pour l'étude.

Exemple : Si une entreprise veut analyser le comportement d'achat de ses clients, au lieu d'étudier toutes les transactions, elle peut sélectionner un échantillon représentatif de transactions pour représenter l'ensemble des clients.

III- Analyse des besoins et sélection des données appropriées

➔ Estimation et Inférence

L'estimation consiste à utiliser des statistiques descriptives pour résumer les caractéristiques d'un échantillon. L'inférence va au-delà, utilisant les résultats de l'échantillon pour tirer des conclusions sur la population entière. Les techniques d'inférence comprennent l'estimation ponctuelle, les intervalles de confiance et les tests d'hypothèses.

- **Estimation ponctuelle :** Utiliser une valeur unique comme meilleure estimation d'un paramètre de population.
- **Intervalles de confiance :** Fournir une plage de valeurs dans laquelle on peut s'attendre à ce que le paramètre de population se situe.
- **Tests d'hypothèses :** Évaluer si les observations de l'échantillon supportent une hypothèse spécifique concernant la population.

Exemple : Une analyse des transactions échantillonnées peut permettre d'estimer la moyenne des dépenses des clients et d'inférer cette information pour toute la population des clients.

III- Analyse des besoins et sélection des données appropriées

➔ Nettoyage et Filtrage des Données

Le nettoyage des données est une étape critique qui consiste à détecter et corriger les erreurs dans les données. Cela inclut la suppression des doublons, la gestion des valeurs manquantes, et la correction des erreurs de saisie.

- Suppression des doublons : Identifier et retirer les enregistrements en double pour éviter de fausser les analyses.
- Gestion des valeurs manquantes : Traiter les données manquantes par suppression ou imputation (remplacement par la moyenne, la médiane, ou des valeurs prédites).
- Correction des erreurs : Identifier et corriger les erreurs dans les données, comme les fautes de frappe ou les incohérences.

Exemple : Lors de l'analyse des transactions des clients, il est important de nettoyer les données pour supprimer les doublons, imputer les valeurs manquantes, et corriger les erreurs de saisie, afin que les résultats de l'analyse soient précis et fiables.

IV- Exploration des caractéristiques des données

L'exploration des données permet de comprendre leur distribution et leurs relations en appliquant différentes méthodes d'analyse.

➔ Analyse unidimensionnelle

L'analyse unidimensionnelle concerne l'étude d'une seule variable à la fois. Elle permet de comprendre sa distribution et ses propriétés statistiques :

- **Moyenne** : Mesure de tendance centrale représentant la valeur moyenne des observations.
- **Médiane** : Valeur qui sépare la moitié inférieure et supérieure des données.
- **Écart-type** : Indicateur de dispersion qui mesure la variabilité des données.
- **Histogrammes** : Représentation graphique montrant la fréquence des valeurs d'une variable.

Exemple : En analysant les revenus mensuels des clients d'une banque, on peut observer une moyenne de 3000 € avec un écart-type de 500 €, indiquant une variation modérée.

IV- Exploration des caractéristiques des données

Analyse multidimensionnelle

L’analyse multidimensionnelle permet d’examiner les relations entre plusieurs variables simultanément :

- **Corrélations** : Mesure de la relation entre deux variables (positive, négative ou nulle).
- **Matrice de dispersion (scatter plot matrix)** : Graphiques permettant d’analyser les interactions entre plusieurs variables quantitatives.

Exemple : Une entreprise analyse la corrélation entre le prix des produits et leur volume de vente pour ajuster sa politique tarifaire.

IV- Exploration des caractéristiques des données

Séries temporelles

Les séries temporelles analysent l'évolution d'une variable sur une période donnée :

- **Détection de tendances** : Observation des augmentations ou diminutions sur le long terme.
- **Saisonnalité** : Identification des variations périodiques (hebdomadaires, mensuelles, annuelles).

Exemple : En analysant les ventes de billets d'avion sur plusieurs années, on remarque une hausse significative pendant les vacances d'été et les fêtes de fin d'année.

IV- Exploration des caractéristiques des données

Probabilités et distributions

L’analyse probabiliste permet de modéliser la distribution des données et de prévoir des comportements futurs :

- **Distribution normale** : Courbe en cloche où la majorité des valeurs sont proches de la moyenne.
- **Distribution binomiale** : Utilisée pour des événements ayant deux issues possibles (succès/échec).
- **Loi de Poisson** : Modélise des événements rares dans un laps de temps donné.

Exemple : Une entreprise de logistique utilise la loi de Poisson pour estimer le nombre de livraisons en retard par jour.

V- Analyse avancée des données

L’analyse avancée permet de réduire la dimensionnalité et de mieux comprendre les structures sous-jacentes.

- **Analyse en Composantes Principales (ACP)** : Réduction de dimension sans perte d’information critique.

Exemple : Une entreprise analysant les performances de ses produits utilise l’ACP pour identifier les facteurs les plus influents sur les ventes.

- **Analyse des Correspondances** : Classification des données qualitatives (simple et multiple).

  Exemple : Une étude de marché catégorise les préférences des consommateurs en fonction de leur tranche d’âge.

- **Clustering** : Identification de groupes homogènes (K-means, hiérarchique).Exemple d’un supermarché qui segmente ses clients en groupes selon leurs habitudes d’achat.

VI- Détection des valeurs atypiques

Les valeurs atypiques (outliers) peuvent fausser l'analyse et doivent être traitées.

- **Détection des valeurs aberrantes** : Méthodes des boîtes à moustaches, z-score, IQR.
- **Traitement des valeurs manquantes** : Suppression, imputation (moyenne, médiane, régression).

Exemple : Une entreprise détecte des prix anormalement élevés dus à des erreurs de saisie.

VII- Interprétation des résultats

L'interprétation des résultats est une étape cruciale de l'analyse des données, car elle permet d'extraire des conclusions exploitables.

- **Comparaison avec les attentes** : Les résultats sont-ils cohérents avec les hypothèses de départ ?
- **Mise en évidence des insights** : Identifier les tendances et patterns importants.
- **Prise de décision** : Traduire les analyses en recommandations concrètes.

Exemple : Une banque utilise l'analyse exploratoire pour détecter les profils à risque de fraude.

VIII- Documentation

Une bonne documentation facilite la reproductibilité et la compréhension des analyses.

- **Rapport d'analyse** : Présentation des résultats sous forme de texte et de graphiques.
- **Visualisation interactive** : Utilisation de notebooks (Jupyter, Google Colab).
- **Traçabilité des étapes** : Stockage des scripts et des jeux de données transformés.

Exemple : Un data scientist documente son processus d'analyse pour permettre à son équipe de le reproduire facilement.

IX- Conclusion

L'analyse exploratoire des données est essentielle pour comprendre les données, identifier des tendances et anticiper des problèmes potentiels. Elle constitue la base de toute analyse avancée et influence directement la qualité des modèles prédictifs ou décisionnels. En maîtrisant ces techniques, les analystes peuvent optimiser la valeur des données et prendre des décisions plus éclairées.