Récap Analyse Exploratoire des Données

Introduction

L'analyse exploratoire des données (AED), c'est un peu comme une première rencontre avec vos données. Imaginez que vous arrivez dans une nouvelle ville : vous allez vous promener, observer les bâtiments, les gens, l'ambiance générale pour vous faire une première idée de l'endroit. C'est la même chose avec l'AED : vous explorez vos données pour comprendre leur structure, identifi er des tendances, des anomalies, et poser les bases de votre analyse plus poussée.

L'AED poursuit plusieurs objectifs clés :

- **Comprendre vos données** : Avant de vous lancer dans des analyses complexes, il est crucial de bien connaître vos données. L'AED vous aide à identifi er les types de variables (quantitatives, qualitatives), leurs distributions, les relations entre elles, etc. C'est comme dresser un portrait-robot de vos données.
- **Détecter des problèmes** : L'AED vous permet de repérer des valeurs aberrantes (outliers), des données manquantes, des erreurs de saisie, qui pourraient fausser vos résultats si elles ne sont pas traitées. C'est un peu comme faire un contrôle technique de vos données avant de prendre la route.
- **Générer des hypothèses** : En explorant vos données, vous pouvez observer des tendances intéressantes, des corrélations surprenantes, qui vous donneront des idées pour des analyses plus approfondies et des modèles prédictifs.
- **Communiquer vos résultats** : L'AED utilise beaucoup de visualisations graphiques, diagrammes) qui sont très utiles pour présenter vos données de manière claire et compréhensible à un public non-technique.

Méthodologie

Source : https://cdn.botpenguin.com/assets/website/Exploratory_Data_Analysis_1_5f8a1c6d39.webp

L'analyse exploratoire des données suit une démarche structurée en trois étapes principales :

Étape 1 : Familiarisation avec les données

En python, la bibliothèque pandas permet de manipuler des données sous forme de Dataframe C'est un peu comme faire connaissance avec une nouvelle personne : on commence par les présentations.

● **Charger les données** : La première étape consiste à importer vos données dans un environnement d'analyse approprié, comme Python avec des bibliothèques comme Pandas ou R.

● **Informations générales** : Ensuite, il est important d'obtenir un aperçu global de vos données : combien d'observations (lignes) et de variables (colonnes) avez-vous ? Quels sont les types de variables (numériques, catégorielles, dates, etc.) ?

● **Premières lignes et statistiques descriptives** : Pour avoir un premier aperçu concret, affichez les premières lignes de votre jeu de données. Calculez également des statistiques descriptives simples pour chaque variable, comme la moyenne, la médiane, l'écart-type, etc.

Cela vous donnera une idée de la distribution et de la variabilité de vos données.

Étape 2 : Visualisation des données

Exemple de visualisation (histogramme)

Une image vaut mille mots, et c'est particulièrement vrai en AED.

● **Visualisations adaptées** : Choisissez les graphiques les plus appropriés pour chaque type de variable. Par exemple, utilisez des histogrammes pour les variables numériques, des diagrammes en barres pour les variables catégorielles, et des nuages de points pour visualiser les relations entre deux variables numériques.

● **Relations entre variables** : Explorez les interactions entre vos variables. Utilisez des tableaux croisés pour les variables catégorielles, des matrices de corrélation pour les variables numériques, ou encore des graphiques plus avancés comme les nuages de points avec couleurs ou tailles variables pour représenter plusieurs dimensions à la fois.

● **Identification de tendances et anomalies** : L'œil humain est très doué pour repérer des motifs, des tendances, ou des points qui sortent de l'ordinaire. Utilisez les visualisations

pour identifi er des tendances intéressantes, des groupes de données, ou des valeurs aberrantes qui méritent une attention particulière.

Étape 3 : Traitement des données

L'*outlier* (valeur aberrante) est un point anormal en comparaison aux autres données . Source

Parfois, vos données ont besoin d'un petit nettoyage avant de pouvoir être analysées en profondeur.

● **Gestion des valeurs manquantes** : Identifi ez les valeurs manquantes dans votre jeu de données, essayez de comprendre pourquoi elles sont manquantes, et décidez de la meilleure façon de les gérer (suppression, imputation par la moyenne ou la médiane, modèles plus complexes).

● **Détection et traitement des valeurs aberrantes (outliers)** : Les outliers sont des valeurs extrêmes qui peuvent biaiser vos analyses. Utilisez des méthodes statistiques ou graphiques pour les détecter, puis décidez s'il faut les supprimer, les corriger, ou les conserver en fonction de leur impact potentiel sur vos résultats.

● **Transformation des données** : Dans certains cas, il peut être utile de transformer vos données pour faciliter l'analyse. Par exemple, vous pouvez normaliser ou standardiser des variables numériques pour les rendre comparables, ou créer de nouvelles variables à partir de variables existantes pour capturer des informations plus pertinentes.

Outils et techniques

L'analyse exploratoire des données s'appuie sur un ensemble d'outils et de techniques qui permettent de décortiquer et de comprendre les données.

Statistiques descriptives

Les statistiques descriptives sont des mesures qui résument les caractéristiques principales de vos données. Elles vous donnent un aperçu rapide de la distribution, de la tendance centrale, et de la dispersion de vos variables.

● **Mesures de tendance centrale** :

○ Moyenne : la somme de toutes les valeurs divisée par le nombre de valeurs.

○ Médiane : la valeur qui sépare les données en deux parties égales.

○ Mode : la valeur la plus fréquente dans les données.

● **Mesures de dispersion** :

○ Écart-type : mesure la dispersion des valeurs autour de la moyenne.

○ Variance : le carré de l'écart-type.

○ Étendue : la diff érence entre la valeur maximale et la valeur minimale.

○ Quartiles : divisent les données en quatre parties égales.

Visualisations

Les visualisations sont des représentations graphiques de vos données qui facilitent leur compréhension et leur interprétation. Elles permettent de repérer des tendances, des motifs, et des anomalies visuellement.

● **Histogrammes** : Représentent la distribution d'une variable numérique en montrant la fréquence de chaque valeur ou intervalle de valeurs.

● **Boîtes à moustaches (box plots)** : Visualisent la distribution d'une variable numérique en montrant les quartiles, la médiane, et les valeurs extrêmes.

● **Nuages de points (scatter plots)** : Représentent la relation entre deux variables numériques en affi chant chaque observation comme un point dans un plan.

● **Diagrammes en barres (bar charts)** : Représentent la fréquence ou la proportion de chaque catégorie d'une variable catégorielle.

● **Camemberts (pie charts)** : Représentent la proportion de chaque catégorie d'une variable catégorielle sous forme de secteurs d'un cercle.

Autres techniques

En plus des statistiques descriptives et des visualisations, d'autres techniques sont couramment utilisées en AED :

● **Tableaux croisés (contingency tables)** : Permettent d'analyser la relation entre deux variables catégorielles en montrant la fréquence de chaque combinaison de catégories.

● **Détection d'outliers** : Utilise des méthodes statistiques ou graphiques pour identifi er les valeurs aberrantes qui s'écartent signifi cativement du reste des données.

● **Imputation de données manquantes** : Consiste à remplacer les valeurs manquantes par des estimations basées sur d'autres informations disponibles, comme la moyenne, la médiane, ou des modèles plus complexes.

L'AED est un domaine riche et en constante évolution, et de nouvelles techniques et outils apparaissent régulièrement. L'important est de choisir les méthodes les plus adaptées à vos données et à vos objectifs d'analyse.

Résumé des points clés

● **L'analyse exploratoire des données (AED)** est une étape cruciale en data science permettant de **comprendre**, **nettoyer** et **extraire des informations** de vos données.

● La **méthodologie** de l'AED comprend trois étapes clés :

○ **Familiarisation** : Charger, examiner la structure et obtenir un aperçu des données.

○ **Visualisation** : Utiliser des graphiques pour identifi er tendances, relations et anomalies

○ **Traitement** : Gérer les valeurs manquantes et aberrantes, transformer les données si besoin

● Les **outils** de l'AED incluent les **statistiques descriptives** (moyenne, médiane...), les **visualisations** (histogrammes, nuages de points...) et d'autres techniques comme les **tableaux croisés** ou l'**imputation**.

● L'AED est **essentielle** pour **guider l'analyse**, **détecter les problèmes** et **prendre des décisions éclairées** basées sur les données.