Voici comment je vais organiser les ressources pour répondre à ta demande :

1. **Un cours théorique** sur **l'analyse exploratoire des données (EDA)**, couvrant :
   * Statistiques descriptives
   * Analyse univariée et bivariée
   * Data visualisation (types de graphiques selon les besoins)
   * Tests de corrélation et interprétation
   * Quelques bonnes pratiques pour préparer les données à la modélisation.
2. **Un Notebook (TP 2)** pour accompagner la théorie avec des exercices pratiques sur un jeu de données fictif ou réel. Ce TP permettra aux étudiants de :
   * Appliquer les statistiques descriptives
   * Visualiser des distributions univariées et bivariées
   * Effectuer des tests de corrélation (Pearson et Spearman)
   * Manipuler les données pour préparer des visualisations pertinentes.

### **🚀 Cours Théorique : Analyse Exploratoire des Données (EDA)**

#### **1. Statistiques descriptives /** Introduction à l'EDA

**Question interactive** : « Pourquoi explorer les données avant toute modélisation ? »

Réponse :

Explorer les données est une étape cruciale avant toute modélisation, car cela permet :

* De comprendre les données : leur structure, les variables présentes, et leurs distributions.
* D'identifier les anomalies : valeurs aberrantes (outliers), données manquantes, doublons.
* D'orienter le choix du modèle : certaines distributions peuvent nécessiter des transformations (par exemple, log-transformation pour les données asymétriques).
* D'éviter les erreurs : travailler sur des données mal préparées peut fausser les résultats du modèle (par exemple, si des outliers ou des erreurs de saisie ne sont pas traités).
* De mieux visualiser les relations entre les variables pour optimiser les choix d’analyse.
* **Types de distributions** :
  + Normale, asymétrique, bimodale : les étudiants peuvent apprendre à les identifier via des graphiques.
* **Objectif** : Résumer les caractéristiques principales d'une variable ou d'un jeu de données.
* **Exemples d'indicateurs** (métriques principales) :
  + Moyenne, médiane, mode
  + Écart-type, variance
  + Min, Max, quartiles (pour identifier les valeurs aberrantes)
  + Distribution des données (histogrammes)

#### **2. Analyse univariée**

* **Objectif** : Étudier chaque variable individuellement.
* **Outils** :
  + Histogrammes
  + Boxplots (pour visualiser les valeurs aberrantes)
  + Countplots (pour les variables catégorielles)
* **Recherche des outliers** :
  + Outil : *Z-score* ou *IQR* (Interquartile Range).

**Question interactive** : « Quelle est la meilleure façon de détecter des outliers dans une variable quantitative ? »

Réponse :

La détection des outliers peut se faire de plusieurs façons :

* Visualisation :
  + Boxplot : Les points en dehors des "whiskers" sont considérés comme des outliers
  + Histogramme ou scatterplot pour observer les valeurs extrêmes.
* Méthodes statistiques :
  + IQR (Interquartile Range) : Calculer l'écart interquartile : Q3 - Q1.Les outliers sont définis comme des valeurs en dehors de [Q1 - 1.5\*IQR, Q3 + 1.5\*IQR].
  + Z-score : On standardise les données. Les observations ayant un Z-score supérieur à 3 (ou inférieur à -3) sont considérées comme des outliers.

**Question interactive :** Quelle est la différence entre le Z-score et l'IQR ?

Réponse :

* Z-score :
  + Définition : Il mesure la distance d'une valeur à la moyenne en termes d'écart-types.
  + Avantage : Fonctionne bien pour les données normalement distribuées.
  + Inconvénient : Sensible aux outliers, car la moyenne et l'écart-type peuvent être biaisés par ces derniers.
* IQR (Interquartile Range) :
  + Définition : Il est basé sur les percentiles (Q1 et Q3) et représente la dispersion des données centrales.
  + Formule = IQR=Q3−Q1.
  + Avantage : Robuste face aux outliers. Il n'utilise pas la moyenne, donc il n'est pas influencé par les valeurs extrêmes.
  + Inconvénient : Moins précis pour les petites tailles d'échantillons.

**Question interactive** : Montre des graphiques (scatterplots, heatmaps) et demande aux étudiants d’interpréter ce qu'ils voient : "Que remarquez-vous ? Quelle conclusion tirez-vous ?"

#### 

#### **3. Analyse bivariée**

* **Objectif** : Étudier les relations entre deux variables.
* **Exemples d'outils** :
  + **Scatterplots** (pour les variables continues)
  + **Boxplots** (Boxplots pour une variable catégorielle vs quantitative]
  + **Heatmaps** pour visualiser des corrélations

#### **4. Tests de corrélation**

* **Objectif** : Identifier les relations linéaires ou monotones entre variables.
  + **Corrélation de Pearson** : Relations linéaires
  + **Corrélation de Spearman** : Relations monotones
* **Visualisation** : Matrice de corrélation avec des heatmaps.

**Question interactive :** "Qu'est-ce que la corrélation de Spearman ? En quoi est-elle différente de Pearson ?"

Réponse :

* Corrélation de Spearman :
  + Elle mesure la corrélation monotone entre deux variables.
  + Elle s’appuie sur les rangs des observations plutôt que sur leurs valeurs brutes.
  + Utilisation : Elle est robuste aux outliers et adaptée aux variables non linéaires.
* Corrélation de Pearson :
  + Elle mesure la corrélation linéaire entre deux variables.
  + Elle suppose une distribution normale et est sensible aux valeurs extrêmes.
* Différence clé :
  + Pearson : relation linéaire entre les variables.
  + Spearman : relation monotone (croissante ou décroissante, pas nécessairement linéaire).

**Question interactive** : Dans quel cas je peux avoir une corrélation de Spearman plus élevée que de Pearson ?

Réponse :

Exemple : Une relation en **forme de courbe croissante** ou de **logarithme**.

**Question interactive** / **Exercice rapide** : « Si deux variables ont une corrélation proche de zéro, cela signifie-t-il qu'elles ne sont pas liées ? Pourquoi ? »

Réponse :

Non, une corrélation proche de zéro ne signifie pas que les variables ne sont pas liées. Cela indique seulement qu'il **n'y a pas de relation linéaire** entre elles.

* Une **relation non linéaire** peut exister même si le coefficient de corrélation est proche de zéro.
* Exemple : Une relation en forme de **U** ou de **parabole** ne sera pas captée par la corrélation de Pearson.

**Question interactive** : "Quelles variables semblent être les plus liées ? Pourquoi ?"

* **Concept de multicolinéarité** :
  + Expliquer pourquoi des variables trop corrélées peuvent poser problème (pratique pour la modélisation).

**Question interactive :** “Qu’est-ce que la multicolinéarité ?“

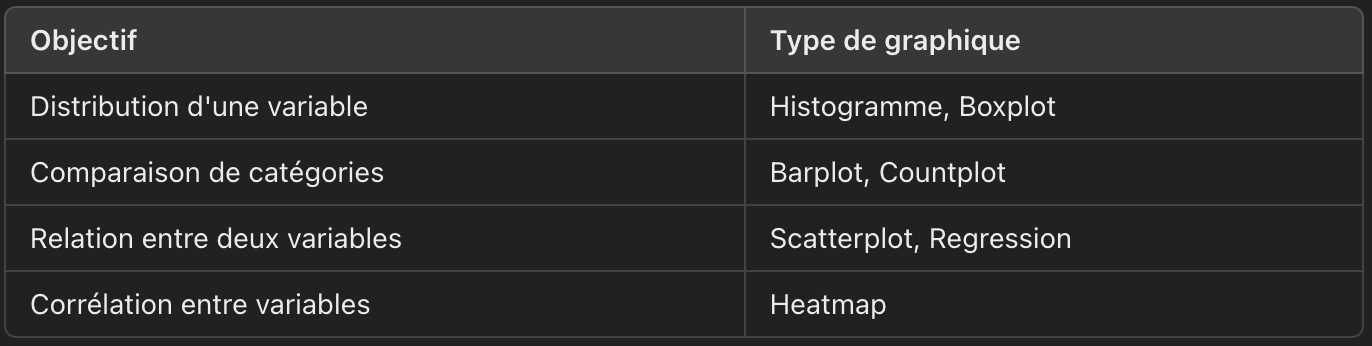
Réponse :

La multicolinéarité désigne une situation dans laquelle deux variables explicatives (ou plus) d'un modèle de régression linéaire sont fortement corrélées entre elles. Cela signifie qu'elles véhiculent une information redondante.

* Lorsque les variables explicatives sont fortement corrélées, il devient difficile pour l'algorithme de déterminer quel poids (ou coefficient) attribuer à chaque variable.
* Les coefficients de régression peuvent changer considérablement avec de petites variations dans les données.
* Concrètement : Si X1 et X2 sont fortement corrélées, l'algorithme ne sait pas s'il doit accorder un poids plus élevé à X1 ou à X2, car leurs effets sont similaires.
* La variance des coefficients augmente avec la multicolinéarité.
  + **XTX** est la matrice de corrélation ou covariance entre les variables explicatives.
  + Si deux variables sont fortement corrélées, X^T . X **quasi-singulière** (non inversible ou proche de l'être).
  + Cela rend l’inversion de X^T X numériquement instable, ce qui fait exploser la variance des coefficients.
* Solutions pour traiter la multicolinéarité :
  + Supprimer une des variables corrélées : Si deux variables véhiculent la même information, supprimez l'une d'entre elles.
  + Combiner les variables : Créez une nouvelle variable (par exemple, la moyenne ou la somme des variables corrélées).
  + Utiliser des méthodes de régularisation : Les modèles comme Ridge Regression (L2) et Lasso Regression (L1) permettent de réduire l'effet de la multicolinéarité en pénalisant les coefficients.

#### **5. Data visualisation**

* Quelques règles pour choisir les graphiques adaptés :



**Question interactive :** "Quel graphique choisiriez-vous pour analyser deux variables quantitatives ?"

**6. Bonnes pratiques**

* Nettoyer les données avant d'analyser (valeurs manquantes, doublons).
* Toujours explorer les données avant de modéliser.
* Utiliser des visualisations adaptées pour comprendre les relations.
  + Réflexion sur l’**interprétabilité des résultats** (un graphique doit parler !)

### **7. Traitement des valeurs manquantes**

### Impact des données manquantes :

1. **Visualisation** : La dispersion semble différente avec des données manquantes.
2. **Corrélation** :
   * La suppression des données **biaisera les résultats**, car les données manquantes peuvent ne pas être aléatoires.
   * Cela peut réduire la puissance de l'analyse (moins de points disponibles).
3. Présenter des techniques de remplissage (imputation) simples.