

## **Projet EDA incluant une variable cible continue et une variable cible discrète :**

Une variable cible, également connue sous le nom de variable dépendante, est la variable dans un ensemble de données qui est prédite ou estimée à partir d'autres variables connues sous le nom de variables indépendantes. Elle est considérée comme la variable la plus importante ou la plus intéressante dans l'analyse de données, car elle est l'objectif de la modélisation ou de la prédiction. Par exemple, dans une étude sur le salaire des employés, la variable cible pourrait être le salaire, tandis que les variables indépendantes pourraient inclure le niveau d'éducation, l'expérience de travail et le secteur d'emploi.

**Introduction :** Dans ce projet, les étudiants seront amenés à réaliser une analyse exploratoire de données (EDA) sur un ensemble de données de leur choix. Le but de ce projet est de donner aux étudiants l'opportunité de mettre en pratique les compétences en matière de manipulation de données, de visualisation et d'analyse de données acquises au cours de leur formation. Le projet devrait également permettre aux étudiants de démontrer leur capacité à communiquer de manière claire et concise les résultats de leur analyse exploratoire.

**Données :** Les étudiants sont libres de choisir n'importe quel ensemble de données pour ce projet, mais il est fortement recommandé de choisir un ensemble de données qui contient un grand nombre de variables pour permettre une analyse plus approfondie. L'ensemble de données doit contenir au moins une variable cible continue et une variable cible discrète.

**Objectifs :** L'objectif principal de ce projet est d'effectuer une EDA complète des données en explorant la structure, les corrélations et les relations entre les variables. Les étudiants devraient être en mesure de répondre aux questions suivantes :

- Quel est le profil des données ? (nombre de variables, nombre d'observations, type de données, etc.)
- Quelles sont les relations entre les variables et comment peuvent-elles être visualisées ?
- Quels sont les modèles ou les tendances qui se dégagent des données ?
- Y a-t-il des variables qui sont fortement corrélées les unes avec les autres ?
- Y a-t-il des variables qui sont fortement corrélées avec les variables cibles ?
- Comment les variables sont-elles distribuées et comment peuvent-elles être transformées si nécessaire ?
- Comment les variables cibles sont-elles liées aux autres variables ?

- Quelles sont les principales conclusions que l'on peut tirer de l'analyse exploratoire des données ?

**Livrable final** (Notebook): Le livrable final devrait inclure les éléments suivants :

1. *Introduction* : Une brève introduction de l'ensemble de données et de l'objectif de l'analyse exploratoire.
2. *Description des données* : Une description détaillée des données, y compris le nombre de variables et d'observations, le type de données, la signification des variables et des observations, ainsi que les éventuelles limitations des données.
3. *Analyse exploratoire des données* : Une analyse complète des données, y compris :
  - *Descriptif statistique* : Une description des variables continues et discrètes.
  - *Visualisation des données* : Des graphiques et des tableaux pour explorer les relations entre les variables et les tendances dans les données.
  - *Corrélation* : Une analyse de la corrélation entre les variables, en mettant l'accent sur les variables cibles.
  - *Traitement des données manquantes* : Les étudiants doivent être en mesure de traiter les données manquantes dans le jeu de données, par exemple en utilisant des techniques d'imputation de données pour remplacer les valeurs manquantes.
  - *Élimination des valeurs aberrantes* : Les étudiants doivent être en mesure d'identifier les valeurs aberrantes dans les données et de décider de les supprimer ou de les conserver dans l'analyse.
  - *Conclusions* : Une synthèse des principales conclusions de l'analyse exploratoire.
4. *Bibliographie* : Toute source utilisée pour ce projet doit être citée dans une bibliographie.
6. *Code source* : Les étudiants devront également soumettre leur code source utilisé pour l'analyse exploratoire des données. Le code source doit être bien documenté et facilement compréhensible.

*Présentation du projet* : Les étudiants devront présenter leur projet, en utilisant des graphiques et des tableaux pour illustrer les résultats clés de leur analyse exploratoire. Ils devront également être prêts à répondre aux questions sur leur analyse exploratoire.

**Évaluation** : Les projets seront évalués sur la base des critères suivants :

- Compréhension des données : Les étudiants doivent démontrer une compréhension approfondie des données qu'ils ont choisies pour leur analyse exploratoire.
- Analyse exploratoire : Les étudiants doivent effectuer une analyse exploratoire complète et détaillée des données, en explorant les relations entre les variables, les modèles et les tendances dans les données.
- Présentation : Les étudiants doivent être capables de présenter leurs résultats de manière claire et concise, en utilisant des graphiques et des tableaux pour illustrer les principaux résultats de leur analyse exploratoire.
- Code source : Les étudiants doivent fournir un code source clair, bien documenté et facilement compréhensible pour leur analyse exploratoire.

Ce projet EDA est une excellente occasion pour les étudiants de mettre en pratique les compétences en matière de manipulation de données, de visualisation et d'analyse de données qu'ils ont acquises au cours. En fournissant des instructions détaillées et des critères d'évaluation clairs, les étudiants devraient être en mesure de mener à bien une analyse exploratoire approfondie et de présenter leurs résultats de manière efficace.