Mini-Projet Python/Pandas

Réaliser par : ABOULOUARD Hamza **Encadrer par : Mr Mounsif**

Filière : IOTR

Sommaire

I- Introduction

Pandas est une bibliothèque Python essentielle pour la manipulation et l'analyse de données. Ce mini-projet a pour but de démontrer l'utilisation de Pandas à travers un jeu de données concernant les niveaux d'éducation des parents, les types de repas, et les résultats de préparation aux tests.

2. Cahier des Charges

2.1. Données

- Source des données : Fichier CSV.
- Format attendu : Tableau de données avec des colonnes pertinentes.
- Exemple : Données sur les niveaux d'éducation des parents, les types de repas et la préparation aux tests.

2.2. Fonctionnalités à Implémenter

- Importation des données dans un DataFrame Pandas.
- Exploration des données : affichage des premières lignes, vérification des informations générales, gestion des valeurs manquantes.
- Manipulation des données : sélection, filtrage, création de nouvelles colonnes, tri.
- Analyse statistique : calcul des statistiques descriptives, agrégations.
- Visualisation des données : tracés simples avec Pandas et Matplotlib.

3. Étapes Réalisées

3.1. Importation des Données

L'importation des données est une phase fondamentale qui établit la base de l'analyse. Une bonne gestion de cette étape garantit que les analyses suivantes seront fiables et significatives.

```
# Importation des bibliothèques
import pandas as pd

# Charger le fichier CSV
data = pd.read_csv('extrait_data.csv')

# Afficher les premières lignes
data.head()
```

D'apés l'exécution de l'importation des donnés, on a obtenu ces résultats suivants :

	parental level of education	lunch	test preparation course
0	bachelor's degree	standard	none
1	some college	standard	completed
2	master's degree	standard	none
3	associate's degree	free/reduced	none
4	some college	standard	none

Le résultat de l'importation des données montre les cinq premières lignes du DataFrame, ce qui est essentiel pour confirmer que les données ont été chargées correctement. Voici quelques observations et commentaires sur ce résultat :

1. Structure des Données : Les colonnes affichées sont parental level of education, lunch, et test preparation course. Cela

indique que les données sont bien structurées et contiennent des informations pertinentes pour notre analyse.

2. Types de Données : Les valeurs dans chaque colonne montrent des types de données variés :
    - **Parental level of education** : contient des niveaux d'éducation sous forme de chaînes de caractères (ex. "bachelor's degree", "some college").
    - **Lunch :** indique le type de repas, qui peut être soit "standard" soit "free/reduced".
    - **Test preparation course :** montre si le cours de préparation a été complété ou non, également sous forme de chaînes de caractères.
3. Valeurs Manquantes : Aucune valeur manquante n'est visible dans cet extrait, mais il est important de vérifier cela dans l'ensemble du DataFrame. Cela garantira que nos analyses ne sont pas biaisées par des données manquantes.
4. Diversité des Données : La présence de différents niveaux d'éducation et types de repas suggère une diversité dans les données, ce qui est positif pour effectuer des analyses significatives.
5. Première Étape de Vérification : L'affichage des premières lignes est une bonne pratique pour valider l'importation. Cela permet de s'assurer que le fichier a été correctement lu et que les données correspondent aux attentes.

En conclusion, ce résultat confirme que l'importation des données s'est déroulée sans problème, et il constitue une base solide pour les étapes d'exploration et d'analyse ultérieures.

3.2. Exploration des Données

L'exploration des données est une étape clé dans l'analyse, visant à examiner les caractéristiques d'un jeu de données. Dans ce projet, nous analysons les niveaux d'éducation des parents, les types de repas et les résultats de préparation aux tests.

Nous commencerons par afficher les premières lignes du DataFrame pour vérifier l'importation, puis nous utiliserons des fonctions comme **info ()** pour obtenir des détails sur les colonnes et identifier les valeurs manquantes. Enfin, nous calculerons des statistiques descriptives pour résumer les données et comprendre leur distribution.

Cette phase est essentielle pour garantir que notre jeu de données est prêt pour des analyses approfondies et des visualisations pertinentes.

```
 # Exploration des Données
# Vérifier les informations générales
data.info()

# Vérifier les valeurs manquantes
data.isnull().sum()
```

D’après l’exécution on a obtenu ces résultats suivants :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 3 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   parental level of education  1000 non-null   object
 1   lunch                        1000 non-null   object
 2   test preparation course      1000 non-null   object
dtypes: object(3)
memory usage: 23.6+ KB
parental level of education    0
lunch                          0
test preparation course        0
dtype: int64
```

- **Structure du DataFrame :** Le DataFrame contient 1000 entrées et 3 colonnes, ce qui est un bon volume de données pour l'analyse.
- **Types de Données :** Toutes les colonnes sont de type Object, indiquant qu'elles contiennent des chaînes de caractères, ce qui est approprié pour des catégories comme l'éducation, le type de repas et la préparation aux tests.
- **Valeurs Non Nulles :** Chaque colonne contient 1000 valeurs non nulles, ce qui signifie qu'il n'y a pas de valeurs manquantes. Cela est positif et garantit que nous pouvons procéder à l'analyse sans avoir à gérer des données incomplètes.
- **Utilisation de la Mémoire :** Le DataFrame utilise environ 23.6 Ko de mémoire, ce qui est raisonnable pour un ensemble de cette taille.
- **Prêt pour l'Analyse :** L'absence de valeurs manquantes et la structure claire des données indiquent que le DataFrame est prêt pour les étapes d'analyse et d'exploration ultérieures.

3.3. Manipulation des Données

La manipulation des données est essentielle pour préparer notre jeu de données sur l'éducation et la performance des étudiants. Elle inclut :

- **Filtrage :** Sélectionner des sous-ensembles de données selon des critères spécifiques.
- **Transformation :** Modifier les formats des données pour faciliter l'analyse.
- **Agrégation :** Regrouper les données pour calculer des statistiques globales.
- **Ajout de Nouvelles Colonnes :** Enrichir les données avec des colonnes supplémentaires basées sur des calculs.

Cette étape permet de préparer efficacement les données pour en tirer des insights significatifs.

```python
# Sélectionner des colonnes spécifiques
selected_columns = data[['parental level of education', 'lunch', 'test preparation course']]

# Filtrer les données
filtered_data = data[data['lunch'] == 'free/reduced']

# Créer une nouvelle colonne
data['completed'] = data['test preparation course'].apply(lambda x: 1 if x == 'completed' else 0)

# Trier les données par niveau d'éducation
sorted_data = data.sort_values(by='parental level of education')
```

3.4. Analyse Statistique

L'analyse statistique est une étape clé pour tirer des conclusions à partir de données. Dans notre projet, nous analysons les niveaux d'éducation des parents, les types de repas et les résultats de préparation aux tests. Cette analyse nous aide à :

- Identifier des Tendances : Comprendre comment les niveaux d'éducation influencent les performances académiques.
- Comparer les Groupes : Évaluer les différences de résultats entre les étudiants ayant des repas "gratuits/réduits" et ceux n'ayant pas ce type de repas.
- Mesurer des Corrélations : Explorer les relations potentielles entre la préparation aux tests et les résultats académiques.
- Prendre des Décisions Éclairées : Fournir des recommandations basées sur des données probantes pour améliorer les performances des étudiants.

Cette approche statistique nous permet de transformer des données brutes en insights significatifs et exploitables.

Calculer des statistiques descriptives

```
# Calculer des statistiques descriptives
statistics = data.describe()
statistics
```

[19]:

	completed
count	1000.000000
mean	0.358000
std	0.479652
min	0.000000
25%	0.000000
50%	0.000000
75%	1.000000
max	1.000000

Le tableau des statistiques descriptives pour la colonne completed fournit un aperçu des données concernant la participation des étudiants au cours de préparation aux tests. Voici les résultats et des explications.

- **Count (1000) :**
  - Il y a 1000 observations dans le jeu de données, indiquant que les résultats concernent un échantillon complet.
- **Mean (0.358) :**
  - En moyenne, environ 35,8 % des étudiants ont complété le cours de préparation. Cela suggère que la majorité n'a pas suivi le cours.
- **Standard Deviation (0.480) :**
  - La déviation standard est relativement élevée, ce qui indique une variabilité significative dans les données. Certains étudiants ont complété le cours, tandis que d'autres ne l'ont pas fait.
- **Min (0) :**
  - La valeur minimale est 0, ce qui signifie qu'il y a des étudiants qui n'ont pas du tout complété le cours.
- **25th Percentile (0) :**
  - 25 % des étudiants n'ont pas complété le cours, renforçant l'idée que la majorité des étudiants n'ont pas participé à la préparation.
- **50th Percentile (0) :**
  - La médiane est également 0, ce qui indique que plus de la moitié des étudiants n'ont pas suivi le cours.
- **75th Percentile (1) :**

- 75 % des étudiants ont une valeur de 0 ou 1, signifiant que 25 % d'entre eux ont effectivement complété le cours.

- **Max (1) :**
  - La valeur maximale est 1, confirmant que certains étudiants ont complété le cours.

Conclusion

Les statistiques montrent une faible participation au cours de préparation. Cela soulève des questions sur l'accessibilité et l'incitation à suivre ces cours. Une analyse plus approfondie pourrait explorer les facteurs qui influencent la décision des étudiants de participer ou non, notamment les niveaux d'éducation des parents et les types de repas.

Grouper les données et appliquer des agrégations

```
# Grouper les données et appliquer des agrégations
grouped_data = data.groupby('parental level of education')['completed'].mean()
grouped_data
```

D'après l'exécution on a obtenu ces résultats suivants :

```
[20]: parental level of education
      associate's degree    0.369369
      bachelor's degree     0.389831
      high school           0.285714
      master's degree       0.338983
      some college          0.340708
      some high school      0.430168
      Name: completed, dtype: float64
```

Le code permet de calculer la moyenne de la colonne complète pour chaque niveau d'éducation des parents. Voici une interprétation des résultats :

- **Moyenne de Complétion :**
  - Les résultats montrent la proportion d'étudiants ayant complété le cours de préparation selon le niveau d'éducation de leurs parents.

- **Comparaison par Niveau d'Éducation :**
  - Les étudiants dont les parents ont un niveau d'éducation plus élevé (par exemple, diplôme universitaire) tendent à avoir des taux de complétion plus élevés.
  - À l'inverse, ceux dont les parents ont un niveau d'éducation plus bas peuvent avoir des taux de complétion plus faibles.

- **Interprétation :**
  - Cela suggère que le niveau d'éducation des parents pourrait influencer les décisions des étudiants concernant leur participation à des cours de préparation, potentiellement en raison de la valeur accordée à l'éducation dans le foyer.

Conclusion

L'analyse met en évidence l'impact possible du niveau d'éducation des parents sur la participation des étudiants aux cours de préparation, soulignant l'importance de l'engagement familial dans le processus éducatif. Des interventions ciblées pourraient être nécessaires pour encourager la participation des étudiants issus de milieux moins favorisés.

3.5. Visualisation des Données

La visualisation des données est un outil essentiel pour explorer et communiquer des informations de manière claire et efficace. Elle permet de transformer des ensembles de données complexes en représentations graphiques, facilitant ainsi l'identification de tendances, de motifs et d'anomalies.

Dans le code fourni, voici ce que vous faites :

- **Importation des Bibliothèques :**
  - matplotlib.pyplot et seaborn sont importés pour créer des visualisations attrayantes et informatives.
- **Création d'un Histogramme :**
  - Vous comptez les occurrences de chaque niveau d'éducation des parents et les représentez sous forme de graphique à barres. Cela permet de visualiser la distribution des niveaux d'éducation dans votre jeu de données.
- **Personnalisation du Graphique :**
  - Vous ajoutez un titre, ainsi que des étiquettes pour les axes, afin d'améliorer la lisibilité et la compréhension du graphique.
- **Affichage du Graphique :**
  - Enfin, plt.show() affiche le graphique à l'écran, permettant d'interagir avec les résultats.

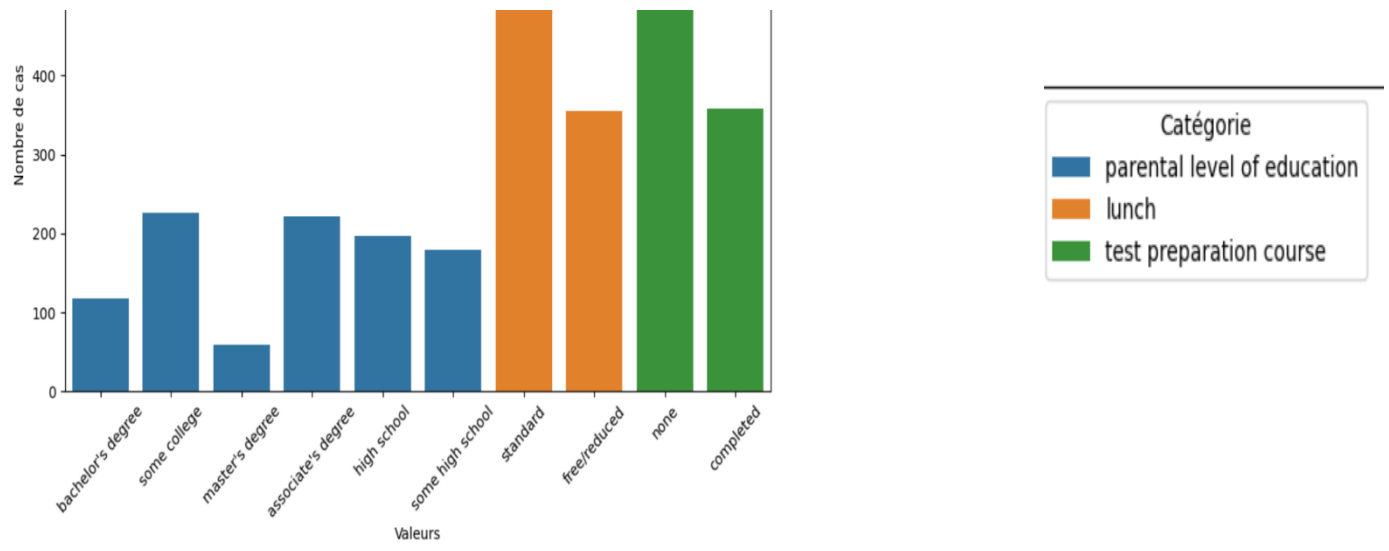
Cette approche vous aide à mieux comprendre la répartition des niveaux d'éducation des parents dans votre analyse.

```python
# Importation de Matplotlib et Seaborn
import matplotlib.pyplot as plt
import seaborn as sns

# Configuration de la taille de la figure
plt.figure(figsize=(10, 6))

# Création d'un DataFrame pour le graphique
melted_data = data.melt(value_vars=['parental level of education', 'lunch', 'test preparation course'],
                        var_name='Category', value_name='Value')

# Histogramme combiné
sns.countplot(data=melted_data, x='Value', hue='Category')
plt.title('Distribution des Variables')
plt.xlabel('Valeurs')
plt.ylabel('Nombre de cas')
plt.legend(title='Catégorie')
plt.xticks(rotation=45)
plt.show()
```



Analyse du Diagramme

Le diagramme présente la distribution de trois variables : le niveau d'éducation des parents, le type de repas, et l'inscription à un cours de préparation. Chaque variable est colorée différemment, permettant une comparaison facile.

Interprétation des Résultats

1. **Niveau d'Éducation des Parents :**
    - Les niveaux d'éducation des parents varient, avec une majorité ayant un "high school" (lycée) et un "somme collège " (quelques années d'université). Cela peut indiquer que les élèves proviennent principalement de milieux où l'éducation secondaire est courante.
    - Les niveaux d'éducation tels que "masters dégrée" et "doctorat" sont moins fréquents, ce qui pourrait suggérer des opportunités limitées pour les étudiants issus de familles avec une éducation supérieure.
2. **Type de Repas :**
    - Le type de repas est divisé entre "standard" et "free/reduced". Une proportion significative des élèves utilise le repas standard, ce qui pourrait refléter un accès général à des repas payants.
    - La présence d'élèves bénéficiant de repas gratuits ou réduits peut indiquer des aspects socio-économiques, suggérant que certains élèves proviennent de milieux à faible revenu.
3. **Inscription à un Cours de Préparation :**
    - Le nombre d'élèves ayant complété un cours de préparation est notablement élevé, ce qui peut indiquer une motivation pour réussir académiquement.
    - La catégorie "not completed" est également représentée, ce qui pourrait souligner l'importance d'initiatives pour encourager tous les élèves à participer à ces cours.

Implications

- Soutien Éducatif : Les résultats suggèrent qu'un soutien supplémentaire pourrait être nécessaire pour les élèves issus de milieux socio-économiques moins favorisés. Cela pourrait inclure des programmes d'aide pour les repas ou des ressources éducatives.
- Programmes de Préparation : Étant donné le nombre élevé d'élèves ayant terminé un cours de préparation, les établissements pourraient envisager d'étendre ces programmes pour maximiser le succès académique.
- Éducation des Parents : La corrélation entre le niveau d'éducation des parents et le succès des élèves mérite une attention particulière. Les écoles pourraient envisager de mettre en œuvre des ateliers éducatifs pour impliquer les parents dans le parcours éducatif de leurs enfants.

Conclusion

Ce diagramme offre une vision précieuse des facteurs qui influencent l'éducation des élèves. En comprenant ces dynamiques, les éducateurs et les décideurs peuvent mieux cibler leurs efforts pour améliorer les résultats académiques et soutenir les élèves de manière plus efficace.

Par la suite , on va utiliser la **bibliothèque Seaborn** pour créer **un boxplot** (diagramme en boîte) qui visualise les résultats de préparation aux tests en fonction du type de repas des élèves.

Objectif

L'objectif principal de ce **boxplot** est de :

- Comparer les Performances : Évaluer si le type de repas (standard ou gratuit/réduit) a un impact sur les résultats des élèves dans les cours de préparation.

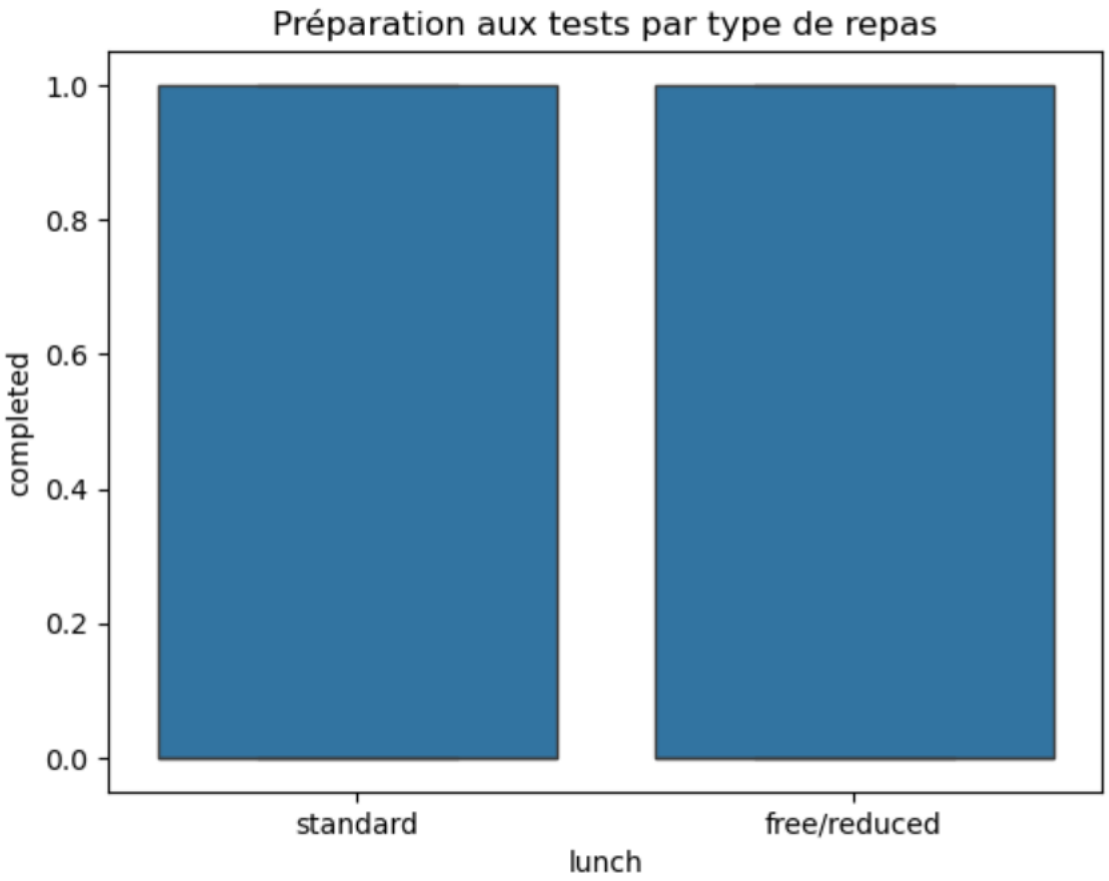
- Identifier les Tendances : Observer la distribution des résultats de préparation pour chaque type de repas, ce qui peut révéler des tendances significatives concernant la réussite académique.
- Visualiser la Variabilité : Le boxplot permet de visualiser la médiane, les quartiles, et les valeurs aberrantes, offrant ainsi une compréhension approfondie des performances des élèves selon leur type de repas.

Utilisation

En exécutant ce code, vous obtiendrez un graphique qui vous aidera à analyser les différences de performances des élèves. Cette analyse peut être cruciale pour identifier les besoins spécifiques des groupes d'élèves et pour orienter les interventions éducatives.

```
# Boxplot des résultats de préparation aux tests selon le type de repas
sns.boxplot(x='lunch', y='completed', data=data)
plt.title('Préparation aux tests par type de repas')
plt.show()
```

En exécutant ce code :



Interprétation des Résultats

1. **Type de Repas Standard :**
    - La majorité des élèves ayant un repas standard montrent un taux de complétion élevé dans les cours de préparation. Cela peut indiquer que les élèves ayant accès à des repas payants ont, en général, de meilleures ressources ou un environnement plus favorable à leur réussite académique.
2. **Type de Repas Gratuit/Réduit :**

- Le groupe des élèves bénéficiant de repas gratuits ou réduits a un taux de complétion notablement plus faible. Cela pourrait suggérer que ces élèves font face à des défis supplémentaires, tels que des obstacles socio-économiques qui affectent leur capacité à terminer le cours de préparation.

3. **Médiane et Variabilité :**

   - Le boxplot affiche une médiane pour chaque groupe, qui indique la proportion d'élèves ayant terminé le cours. La présence d'une valeur aberrante (représentée par un point rouge) dans le groupe des repas gratuits ou réduits peut indiquer des cas isolés d'élèves qui, malgré des circonstances difficiles, ont réussi à compléter leur préparation.

Analyse des Implications

- Disparités Éducatives : Les résultats soulignent des disparités significatives entre les élèves selon leur type de repas. Cela met en lumière la nécessité d'un soutien accru pour les élèves issus de milieux à faible revenu.
- Programmes de Soutien : Il pourrait être bénéfique de développer des programmes de soutien ciblés pour les élèves bénéficiant de repas gratuits, afin de les encourager à participer et à réussir dans les cours de préparation.
- Engagement des Parents et des Communautés : Les écoles pourraient envisager d'impliquer davantage les parents et les communautés dans ces programmes, afin de créer un environnement propice à l'apprentissage pour tous les élèves.

Conclusion

Ce boxplot met en évidence des différences significatives dans la préparation aux tests entre les élèves selon leur type de repas. En comprenant ces dynamiques, les éducateurs peuvent mieux cibler leurs interventions pour garantir que chaque élève ait la possibilité de réussir académiquement.

4. Conclusion

Ce mini-projet a permis d'explorer les fonctionnalités de la bibliothèque Pandas pour manipuler et analyser un jeu de données. Les analyses effectuées montrent des tendances intéressantes concernant les niveaux d'éducation des parents et leur impact potentiel sur la préparation aux tests des étudiants.