

data_challenge

May 17, 2025

```
[1]: from google.colab import drive  
drive.mount('/content/drive')
```

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

```
[2]: import pandas as pd
```

SID44 : data challenge

TABLE DES MATIÈRES

INTRODUCTION

PROBLÉMATIQUE ET OBJECTIFS

EXPLORATION DES DONNÉES

Présentation des jeux de données

Fusion des données

Nettoyage et traitement des valeurs manquantes

ANALYSE EXPLORATOIRE

Analyse univariée

Analyse bivariée

Comparaison régionale

MODÉLISATION NON SUPERVISÉE

K-means & PCA

Interprétation des clusters

Amélioration des performances

RÉSULTATS ET DISCUSSIONS

CONCLUSION ET RECOMMANDATIONS

INTRODUCTION

L'éducation de base constitue un pilier fondamental du développement humain et économique d'un pays. En Mauritanie, malgré une scolarisation primaire relativement élevée, de nombreux défis

persistent quant à la qualité de l’enseignement et aux conditions d’apprentissage, en particulier dans les régions rurales et défavorisées. Le pays affiche un indice de capital humain de seulement 0,38, témoignant d’un accès limité à une éducation efficace et à des services de santé de qualité. De plus, le système éducatif souffre de plusieurs insuffisances : manque d’enseignants qualifiés, taux d’absentéisme élevé, faible niveau pédagogique, infrastructures dégradées, et pénurie de matériels d’enseignement. L’introduction d’un système bilingue (arabe-français) depuis 1999 a également complexifié l’apprentissage, les enseignants comme les élèves rencontrant des difficultés linguistiques dans la maîtrise des deux langues.

Dans ce contexte, le programme “The Future is Ours”, financé par le département américain de l’agriculture (USDA) dans le cadre de l’initiative McGovern-Dole (Food for Education and Child Nutrition), vise à améliorer la sécurité alimentaire, la santé, et surtout la littératie en primaire, en promouvant un environnement scolaire favorable à l’apprentissage. L’extension du programme sous le nom “Bridging the Future” (2022–2027) couvre 111 écoles primaires dans les régions du Brakna, Gorgol, et Tagant, touchant directement plus de 115 000 bénéficiaires : élèves, enseignants, directeurs, parents, et personnels communautaires.

L’intervention se structure autour de cinq axes majeurs : 1. Amélioration de la fréquentation scolaire par la fourniture de repas scolaires quotidiens, 2. Renforcement des pratiques sanitaires et nutritionnelles (eau, hygiène, déparasitage), 3. Réhabilitation des infrastructures scolaires (latrines, points d’eau, cantines), 4. Développement des compétences des enseignants à travers des formations bilingues et pédagogiques adaptées, 5. Implication des communautés locales via les COGES, APE et associations locales pour renforcer la pérennité des actions.

L’enquête de référence conduite en 2023 fournit une base de données exhaustive sur l’état initial des apprentissages, des pratiques pédagogiques, des infrastructures et de l’implication communautaire. Les données collectées auprès de 96 écoles à travers des tests de lecture (EGRA), des questionnaires, observations de classes, entretiens et groupes de discussion, permettent une analyse fine des facteurs influençant la qualité de l’enseignement et l’attention des élèves.

C’est dans cette optique que s’inscrit notre étude, qui vise à répondre à la problématique suivante :

Dans quelle mesure les conditions de travail, la formation des enseignants et les ressources pédagogiques influencent-elles la qualité de l’enseignement et l’attention des élèves dans les écoles bénéficiaires du programme “The Future is Ours” ?

L’analyse présentée permettra d’identifier les leviers prioritaires pour améliorer l’enseignement primaire dans ces régions, tout en proposant des recommandations concrètes pour ajuster le programme à mi-parcours.

PROBLÉMATIQUE ET OBJECTIFS

Dans quelle mesure les conditions de travail, la formation des enseignants et les ressources pédagogiques influencent-elles la qualité de l’enseignement et l’attention des élèves dans les écoles bénéficiaires du programme “The Future is Ours” ?

Cette problématique guide notre analyse des données issues de l’enquête. Elle vise à identifier les principaux leviers d’action pour renforcer la qualité de l’éducation dans les régions ciblées et formuler des recommandations pour un pilotage plus efficace du programme à moyen terme.

EXPLORATION DES DONNÉES

[]:

[5]: !pip install pyreadstat

```
Collecting pyreadstat
  Downloading pyreadstat-1.2.8-cp311-cp311-
manylinux_2_17_x86_64.manylinux2014_x86_64.whl.metadata (1.0 kB)
Requirement already satisfied: pandas>=1.2.0 in /usr/local/lib/python3.11/dist-
packages (from pyreadstat) (2.2.2)
Requirement already satisfied: numpy>=1.23.2 in /usr/local/lib/python3.11/dist-
packages (from pandas>=1.2.0->pyreadstat) (2.0.2)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.11/dist-packages (from pandas>=1.2.0->pyreadstat)
(2.9.0.post0)
Requirement already satisfied: pytz>=2020.1 in /usr/local/lib/python3.11/dist-
packages (from pandas>=1.2.0->pyreadstat) (2025.2)
Requirement already satisfied: tzdata>=2022.7 in /usr/local/lib/python3.11/dist-
packages (from pandas>=1.2.0->pyreadstat) (2025.2)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.11/dist-
packages (from python-dateutil>=2.8.2->pandas>=1.2.0->pyreadstat) (1.17.0)
Downloading
pyreadstat-1.2.8-cp311-cp311-manylinux_2_17_x86_64.manylinux2014_x86_64.whl (2.9
MB)

2.9/2.9 MB
20.5 MB/s eta 0:00:00
Installing collected packages: pyreadstat
Successfully installed pyreadstat-1.2.8
```

[6]: import pyreadstat

[76]: df_1 = pd.read_stata('/content/drive/MyDrive/DATA CHALLENGE/DATA CHALLENGE/
↳Quant data before appending/teacher survey.dta')

```
<ipython-input-76-68c0918e0aca>:1: UnicodeWarning:
One or more strings in the dta file could not be decoded using utf-8, and
so the fallback encoding of latin-1 is being used. This can happen when a file
has been incorrectly encoded by Stata or some other software. You should verify
the string values returned are correct.
df_1 = pd.read_stata('/content/drive/MyDrive/DATA CHALLENGE/DATA
CHALLENGE/Quant data before appending/teacher survey.dta')
```

[77]: df_1, meta_1 = pyreadstat.read_dta('/content/drive/MyDrive/DATA CHALLENGE/DATA_
↳CHALLENGE/Quant data before appending/teacher survey.dta', encoding='latin1')

[78]: meta_1.column_names_to_labels

[78]: {'start': 'Start time',
 'today': 'Today',

'team_lead': 'Name of the team leader',
 'enum_name': 'Name of enumerator',
 'gps': 'GPS',
 '_v1': '_gps_latitude',
 '_v2': '_gps_longitude',
 '_v3': '_gps_altitude',
 '_v4': '_gps_precision',
 'region': 'Select the region',
 'department': 'Select the department',
 'commune': 'Select the commune',
 'school': 'Select the school',
 'consent': 'Hello! How are you? I am speaking with you today on behalf of the
 Future is Ours',
 'gender': 'Gender of teacher',
 'grades': 'What grades do you teach?',
 '_v5': 'grades/First grade',
 '_v6': 'grades/Second grade',
 '_v7': 'grades/Third grade',
 '_v8': 'grades/Other',
 'grades_other': 'If other, please specify',
 '_v9': 'grades_other/Fourth grade',
 '_v10': 'grades_other/Fifth grade',
 '_v11': 'grades_other/Sixth grade',
 'language': 'What languages do you teach in school?',
 '_v12': 'language/Arabic',
 '_v13': 'language/French',
 '_v14': 'language/Other (Specify)',
 'language_other': 'If other, please specify',
 'language_home': 'What language do you speak most of the time at home?',
 'language_home_other': 'If other, please specify',
 'attent1': 'On average, since the beginning of this school year, how many times
 do YOU inter',
 'attent2': 'Since the beginning of the school year, do you feel student
 attention has declin',
 'attent3': 'Why do you feel student attention has declined during your
 lessons?',
 '_v15': 'attent3/Fatigue or tiredness',
 '_v16': 'attent3/Malnourishment',
 '_v17': 'attent3/Lack of interest in the topic',
 '_v18': 'attent3/External distractions (e.g., noise, technology)',
 '_v19': 'attent3/Personal issues affecting student focus',
 '_v20': 'attent3/Difficulty understanding the material',
 '_v21': 'attent3/Classroom disruptions or behavior issues',
 '_v22': 'attent3/Other (please specify)',
 'attent3_other': 'If other, please specify',
 'train1': 'On a scale of 1 (not well trained) to 5 (very well trained), to what
 extent do Y',

```

'train_reading': 'When did you last receive training related to teaching
reading?',
'train_reading_who': 'Who provided this training?',
'educ_mat': 'Since the start of the 2020 school year, have you received new
educational mater',
'educ_mat_what': 'What materials did you receive?',
'educ_mat_needs': 'On a scale of 1 (does not meet my needs) to 5 (exceeds my
needs), to what extent',
'educ_mat_explain': 'Why did you give your teaching materials this score?',
'absent': 'During this term, how many days of school were you absent?',
'absent_reason': 'What was the main reason for your absence?',
'absent_reason_other': 'If other, please specify',
'attend': 'Do you take attendance in your class every day?',
'attend_commun': 'How often do you communicate attendance data to the school
principal?',
'attend_commun_other': 'If other, please specify',
'program': 'Which of the following aspects of The Future is Ours project has
your school rec',
'_v23': 'program/Providing food for school meals',
'_v24': 'program/Deworming medicine',
'_v25': 'program/Training on food preparation, storage, sanitation, and
hygiene',
'_v26': 'program/Provision of cooking equipment',
'_v27': 'program/Parent training on school participation',
'_v28': 'program/New Arabic reading materials',
'_v29': 'program/Teacher training',
'_v30': 'program/Reading clubs',
'_v31': 'program/Other',
'program_other': 'If other, please specify',
'add_info': 'Is there anything else you would like to tell me that I have not
asked about?',
'add_info2': 'FOR ENUMERATOR: Please note any additional information that is
important for us',
'instanceID': 'instanceID',
'instanceName': 'instanceName',
'_v32': '_id',
'_v33': '_uuid',
'_v34': '_submission_time'}

```

Présentation des jeux de professeur

```
[79]: df_1
```

```

[79]:
      start      today  team_lead  enum_name  \
0  2024-05-02T12:57:06.484Z  2.030227e+12      3      304
1  2024-05-02T13:24:28.419Z  2.030227e+12      3      301
2  2024-04-22T11:23:26.006Z  2.029363e+12      5      502
3  2024-04-22T12:32:23.747Z  2.029363e+12      5      502

```

4	2024-04-22T11:37:32.859Z	2.029363e+12	5	502
..
429	2024-04-26T09:57:20.434Z	2.029709e+12	4	402
430	2024-04-26T10:14:29.202Z	2.029709e+12	4	402
431	2024-04-25T13:52:01.775Z	2.029622e+12	4	402
432	2024-04-25T12:59:14.150Z	2.029622e+12	4	402
433	2024-04-25T14:11:59.673Z	2.029622e+12	4	402

			gps		_v1	_v2	_v3	_v4	\
0	17.3502652	-13.8821416	83.0	4.983	17.350265	-13.882142	83.0	4.983	
1	17.350582	-13.882404	60.8	4.8	17.350582	-13.882404	60.8	4.800	
2	17.1218787	-14.0668247	62.8	3.6	17.121879	-14.066825	62.8	3.600	
3	17.1219733	-14.0666367	65.5	4.5	17.121973	-14.066637	65.5	4.500	
4	17.1218681	-14.0667606	65.5	4.65	17.121868	-14.066761	65.5	4.650	
..	
429	16.4063057	-13.1360057	78.3	4.72	16.406306	-13.136006	78.3	4.720	
430	16.4068884	-13.1361643	76.4	4.242	16.406888	-13.136164	76.4	4.242	
431	16.1580684	-12.722907	64.2	4.666	16.158068	-12.722907	64.2	4.666	
432	16.1580627	-12.7228578	60.3	4.9	16.158063	-12.722858	60.3	4.900	
433	16.158081	-12.7229026	35.1	4.516	16.158081	-12.722903	35.1	4.516	

	region	...	_v30	_v31	program_other	\
0	1	...	0	0		
1	1	...	0	0		
2	1	...	0	0		
3	1	...	0	0		
4	1	...	0	0		
..	
429	2	...	1	0		
430	2	...	1	0		
431	2	...	1	0		
432	2	...	1	1	Donner des CSB aux femmes enceintes	
433	2	...	1	0		

	add_info	\
0	Je voudrais que les responsables du projet l'a...	
1	- besoin de livres \n- besoin de rÃ©fectoire	
2	Selon votre programme pourquoi y'as pas une fo...	
3	Je vous remercie d'avoir passer nous voir	
4	Est vous Ãªtes bien de ce tÃ¢che que vous faites	
..	...	
429	C'est un bon projet mais nous aimerons que cou...	
430	C'est un bon projet qui fait bÃ©nÃ©ficier Ã c...	
431	On aimerait faire du jardinage scolaire pour i...	
432	Nous voulons que vous aidiez nos cuisiniÃ¨res,...	
433	Nous aimerons avoir des ordinateurs pour trava...	

```

                                add_info2 \
0   Je remarque qu'il un probl me de s curit  d...
1   L'enseignant de la 1ere ann e est le directeur
2   Est ce que vos  l ves ont l'habitude de baga...
3   Pourquoi on  volue pas les  l ves de 1 re ...
4   Depuis quant vous enseigner ici
..
429 La salle est petite et en zinc,certaines table...
430 Le tableau n'est pas bon,les tables en b ton,...
431                                     Rien
432                                     Rien
433                                     Rien

```

```

                                instanceID      instanceName      _v32 \
0   uuid:d1f086a2-330f-4999-81c1-882b8a05ab9f      1 - 2024-05-02      136299398
1   uuid:220fbbf1-aef0-444e-8aab-7d9ca4ffb9ab      1 - 2024-05-02      136299340
2   uuid:094f2a45-8e91-4102-a031-6c4329f04676      2 - 2024-04-22      135786199
3   uuid:86e49a61-3c14-41d5-8ed0-9621f1e0bf6f      2 - 2024-04-22      135786201
4   uuid:240e3d8e-aea4-4839-8e20-95ab4f574517      2 - 2024-04-22      135786200
..
429 uuid:6dd14b29-6944-4692-9d78-3a8dfa6e7b54      145 - 2024-04-26      135942122
430 uuid:41e1df2a-6b79-4e84-b274-10ec5ca5b291      145 - 2024-04-26      135942124
431 uuid:1251140f-cd30-435a-81da-87e3cb28062c      146 - 2024-04-25      135909380
432 uuid:37d39fe1-3a08-4b01-aa74-c1c77ac622b5      146 - 2024-04-25      135909378
433 uuid:83c88850-8938-41ab-a170-b06bd89f9050      146 - 2024-04-25      135909385

```

```

                                _v33      _v34
0   d1f086a2-330f-4999-81c1-882b8a05ab9f      2.030281e+12
1   220fbbf1-aef0-444e-8aab-7d9ca4ffb9ab      2.030281e+12
2   094f2a45-8e91-4102-a031-6c4329f04676      2.029477e+12
3   86e49a61-3c14-41d5-8ed0-9621f1e0bf6f      2.029477e+12
4   240e3d8e-aea4-4839-8e20-95ab4f574517      2.029477e+12
..
429 6dd14b29-6944-4692-9d78-3a8dfa6e7b54      2.029766e+12
430 41e1df2a-6b79-4e84-b274-10ec5ca5b291      2.029766e+12
431 1251140f-cd30-435a-81da-87e3cb28062c      2.029680e+12
432 37d39fe1-3a08-4b01-aa74-c1c77ac622b5      2.029680e+12
433 83c88850-8938-41ab-a170-b06bd89f9050      2.029680e+12

```

[434 rows x 74 columns]

[]:

Pr sentation des variables des etudiant

```

[73]: df_etudiant = pd.read_stata('/content/drive/MyDrive/DATA CHALLENGE/DATA_
      ↪CHALLENGE/Quant data before appending/egra and student survey.dta')

```

```
[75]: df_etudiant, meta_2 = pyreadstat.read_dta('/content/drive/MyDrive/DATA_
↳CHALLENGE/DATA CHALLENGE/Quant data before appending/egra and student survey.
↳dta', encoding='latin1')
meta_2.column_names_to_labels
```

```
[75]: {'start': 'Start time',
'today': 'Today',
'team_lead': 'Name of the team leader',
'enum_name': 'Name of enumerator',
'gps': 'GPS',
'_v1': '_gps_latitude',
'_v2': '_gps_longitude',
'_v3': '_gps_altitude',
'_v4': '_gps_precision',
'region': 'Select the region',
'department': 'Select the department',
'commune': 'Select the commune',
'school': 'Select the school',
'gender': 'Gender of Student',
'grade_student': 'Select the grade',
'egra_student': 'Does this student need to take the student survey?',
'consent': 'Hello! How are you? I am speaking with you today on behalf of the
Future is Ours',
'egra_random': 'egra_random',
'f1_wrong': 'How many characters did the student read incorrectly?',
'f1_total': 'What was the number of the last character the child read before
time ran out? En',
'f1_end_early': 'Did you need to stop the assessment before one minute because
the child was able',
'f1_end_early_time': 'How many seconds elapsed by the time the student finished
the assessment?',
'f1_confirm': 'Please confirm that the student read ${f1_total}',
'f2_wrong': 'How many syllables did the student read incorrectly?',
'f2_total': 'What was the number of the last syllable the child read before
time ran out? Ent',
'f2_end_early': 'Did you need to stop the assessment before one minute because
the child was able',
'f2_end_early_time': 'How many seconds elapsed by the time the student finished
the assessment?',
'f2_confirm': 'Please confirm that the student read ${f2_total}',
'f3_wrong': 'How many words did the student read incorrectly?',
'f3_total': 'What was the number of the last word the child read before time
ran out? Enter 5',
'f3_end_early': 'Did you need to stop the assessment before one minute because
the child was able',
'f3_end_early_time': 'How many seconds elapsed by the time the student finished
the assessment?',
```


'f3_confirm': 'Please confirm that the student read \${f3_total}',
 'f4_wrong': 'How many words did the student read incorrectly?',
 'f4_total': 'What was the number of the last word the child read before time
 ran out? Enter 2',
 'f4_end_early': 'Did you need to stop the assessment before one minute because
 the child was able',
 'f4_end_early_time': 'How many seconds elapsed by the time the student finished
 the assessment?',
 'f4_confirm': 'Please confirm that the student read \${f4_total}',
 'f5_wrong': 'How many words did the student read incorrectly?',
 'f5_total': 'What was the number of the last word the child read before time
 ran out? Enter 4',
 'f5_end_early': 'Did you need to stop the assessment before one minute because
 the child was able',
 'f5_end_early_time': 'How many seconds elapsed by the time the student finished
 the assessment?',
 'f5_confirm': 'Please confirm that the student read \${f5_total}',
 'f6_q1': 'What does Fatima do when she returns from school?',
 'f6_q2': 'Who sweeps the yard?',
 'f6_q3': 'What is Fatima's friend's name?',
 'f6_q4': 'What does Fatima do with her friend in the afternoon?',
 'ar1_wrong': 'How many characters did the student read incorrectly?',
 'ar1_total': 'What was the number of the last character the child read before
 time ran out? En',
 'ar1_end_early': 'Did you need to stop the assessment before one minute because
 the child was able',
 'ar1_end_early_time': 'How many seconds elapsed by the time the student
 finished the assessment?',
 'ar1_confirm': 'Please confirm that the student read \${ar1_total}',
 'ar2_wrong': 'How many syllables did the student read incorrectly?',
 'ar2_total': 'What was the number of the last syllable the child read before
 time ran out? Ent',
 'ar2_end_early': 'Did you need to stop the assessment before one minute because
 the child was able',
 'ar2_end_early_time': 'How many seconds elapsed by the time the student
 finished the assessment?',
 'ar2_confirm': 'Please confirm that the student read \${ar2_total}',
 'ar3_wrong': 'How many words did the student read incorrectly?',
 'ar3_total': 'What was the number of the last word the child read before time
 ran out? Enter 5',
 'ar3_end_early': 'Did you need to stop the assessment before one minute because
 the child was able',
 'ar3_end_early_time': 'How many seconds elapsed by the time the student
 finished the assessment?',
 'ar3_confirm': 'Please confirm that the student read \${ar3_total}',
 'ar4_wrong': 'How many words did the student read incorrectly?',
 'ar4_total': 'What was the number of the last word the child read before time

ran out? Enter 2',
 'ar4_end_early': 'Did you need to stop the assessment before one minute because the child was able',
 'ar4_end_early_time': 'How many seconds elapsed by the time the student finished the assessment?',
 'ar4_confirm': 'Please confirm that the student read \${ar3_total}',
 'ar5_wrong': 'How many words did the student read incorrectly?',
 'ar5_total': 'What was the number of the last word the child read before time ran out? Enter 4',
 'ar5_end_early': 'Did you need to stop the assessment before one minute because the child was able',
 'ar5_end_early_time': 'How many seconds elapsed by the time the student finished the assessment?',
 'ar5_confirm': 'Please confirm that the student read {ar5_total}',
 'ar7_q1': '1. Who commanded you. Is it your mother and father?',
 'ar7_q2': '2. What did they teach you and your siblings?',
 'ar7_q3': '3. What have they planted in your souls?',
 'grade': 'What grade are you in?',
 'age': 'What is your age?',
 'ethnicity': 'What is your ethnicity?',
 'ethnicity_other': 'Specify Other Ethnicity',
 'home_language': 'What languages do you most often use when you are at home?',
 '_v5': 'home_language/French',
 '_v6': 'home_language/Hassanya',
 '_v7': 'home_language/Pulaar',
 '_v8': 'home_language/Wolof',
 '_v9': 'home_language/Soninke',
 '_v10': 'home_language/Other',
 'home_language_other': 'Specify other language',
 'home_own_rent': 'Does your family own or rent a house?',
 'econ_hungry': 'Gone to sleep at night feeling hungry?',
 'econ_water': 'Gone without enough clean water for the home to use?',
 'econ_med': 'Gone without needed medicine or medical treatment?',
 'econ_cash': 'Gone without cash income?',
 'attentiveness1': 'When you are at school, do you feel like you can pay attention to your teacher?',
 'attentiveness2': 'When do you think you can be more attentive to your teacher?',
 'school_eat': 'Where do you usually eat when you are at school?',
 'school_eat_other': 'If other, please specify',
 'school_food_like': 'Do you like the food you eat during school?',
 'school_miss': 'Have you ever missed school to eat?',
 'school_miss_reason': 'Why do you miss school to go eat?',
 '_v11': 'school_miss_reason/Arrives late after breakfast at home',
 '_v12': 'school_miss_reason/Comes home for lunch and does not return',
 '_v13': 'school_miss_reason/Other',
 'school_miss_reason_oth': 'If other, please specify',

'school_water': 'Where do you usually get water for drinking during school?',
 'school_water_other': 'If other, please specify',
 'school_water_qual': 'Do you think the water quality is good or not good?',
 'school_water_bad': "Why isn't the water good?",
 '_v14': 'school_water_bad/Tastes bad',
 '_v15': 'school_water_bad/The water is dirty',
 '_v16': 'school_water_bad/The water area is not clean',
 '_v17': 'school_water_bad/Water is not always available',
 '_v18': 'school_water_bad/Other',
 'school_water_bad_oth': 'If other, please specify',
 'school_miss_water': "Do you ever miss school because you don't want to drink the water?",
 'school_latrines': 'Are you comfortable using the latrines at school?',
 'school_latrines_uncomf': 'Why are you uncomfortable using the latrines at school?',
 '_v19': 'school_latrines_uncomf/They are not clean',
 '_v20': 'school_latrines_uncomf/There is not enough privacy',
 '_v21': 'school_latrines_uncomf/There is no place to wash my hands',
 '_v22': 'school_latrines_uncomf/They are not safe',
 '_v23': 'school_latrines_uncomf/There is no water',
 '_v24': 'school_latrines_uncomf/Other',
 'school_latrines_uncomf_oth': 'If other, please specify',
 'school_latrines_no_use': 'Do you ever miss school because you do not want to use the school latrine?',
 'days_missed': 'How many days did you not go to school last semester?',
 'days_missed_reason': 'On the days you missed, what were the main reasons you did not go to school?',
 '_v25': 'days_missed_reason/I was sick',
 '_v26': 'days_missed_reason/Someone in my family was sick',
 '_v27': 'days_missed_reason/I needed to work on the farm',
 '_v28': 'days_missed_reason/I needed to do other work',
 '_v29': 'days_missed_reason/The latrine facilities at the school are not good',
 '_v30': "days_missed_reason/The school wasn't accessible because of rainfall",
 '_v31': 'days_missed_reason/Other',
 'days_missed_reason_other': 'If other, please specify',
 'days_late': 'How many days did you leave school early or come to school late last semester?',
 'days_late_reason': 'On the days you went home from school early or came to school late, what were th',
 '_v32': 'days_late_reason/I was sick',
 '_v33': 'days_late_reason/Someone in my family was sick',
 '_v34': 'days_late_reason/I needed to work on the farm',
 '_v35': 'days_late_reason/I needed to do other work',
 '_v36': 'days_late_reason/I needed to go home to eat',
 '_v37': 'days_late_reason/I needed to get water',
 '_v38': 'days_late_reason/The latrine facilities at school are not good',
 '_v39': 'days_late_reason/Other',

'days_late_reason_other': 'If other, please specify',
 'reading_participate': 'Did you participate in any extracurricular reading activities at the school this',
 'reading_what': 'What kind of extracurricular reading activities did you participate in?',
 '_v40': 'reading_what/Reading club',
 '_v41': 'reading_what/Reading competition',
 '_v42': 'reading_what/Educational day',
 '_v43': 'reading_what/Other',
 'reading_what_other': 'If other, please specify',
 'reading_participate_no': 'If you did not participate any extracurricular activities, why not?',
 'reading_participate_no_other': 'If other, please specify',
 'reading_activity': 'Which of the following reading activities do you do in class?',
 '_v44': 'reading_activity/My teacher reads to us',
 '_v45': 'reading_activity/My teacher helps us to read by ourselves during a lesson',
 '_v46': 'reading_activity/My teacher asks us to read out louds in class',
 '_v47': 'reading_activity/My teacher allows us to do reading activities in small groups',
 '_v48': 'reading_activity/My teacher assigns me reading to do at home',
 '_v49': 'reading_activity/My teacher gives us reading games and puzzles',
 '_v50': 'reading_activity/Other',
 'reading_activity_other': 'If other, please specify',
 'attention_class': 'How easy is it for you to pay attention to your teacher in class?',
 'attention_class_why': 'Why do you find it difficult to pay attention in class?',
 'school_wash': 'How many times a day do you wash your hands at school?',
 'school_wash_why': 'Why don't you wash your hands at school?',
 '_v51': 'school_wash_why/There is no place for washing my hands',
 '_v52': 'school_wash_why/There is no water to wash my hands',
 '_v53': 'school_wash_why/There is no soap',
 '_v54': 'school_wash_why/The facilities are dirty',
 '_v55': 'school_wash_why/Other',
 'school_wash_often_other': 'Other reason for not washing hands at school',
 'school_wash_when': 'When do you typically wash your hands when you are at school?',
 '_v56': 'school_wash_when/Before eating',
 '_v57': 'school_wash_when/Before touching or preparing food',
 '_v58': 'school_wash_when/Before giving food to someone else',
 '_v59': 'school_wash_when/When I have dirt on my hands',
 '_v60': 'school_wash_when/After touching something dirty',
 '_v61': 'school_wash_when/After touching an animal',
 '_v62': 'school_wash_when/After using a latrine',
 '_v63': 'school_wash_when/Other',

```

'school_wash_when_other': 'Other time for handwashing at school',
'school_wash_type': 'What do you typically use to wash your hands at school?',
'_v64': 'school_wash_type/Water',
'_v65': 'school_wash_type/Ashes',
'_v66': 'school_wash_type/Sand',
'_v67': 'school_wash_type/Soap',
'_v68': 'school_wash_type/Other',
'school_wash_type_other': 'Other handwashing material at school',
'home_wash': 'How many times a day do you wash your hands at home?',
'home_wash_why': 'Why don't you usually wash your hands at home?',
'_v69': 'home_wash_why/There is no place for washing my hands at home',
'_v70': 'home_wash_why/There is no water to wash my hands',
'_v71': 'home_wash_why/There is no soap',
'_v72': 'home_wash_why/The facilities are dirty',
'_v73': 'home_wash_why/Other',
'home_wash_why_other': 'Other reason for not washing hands at home',
'home_wash_when': 'When do you typically wash your hands when you are at
home?',
'_v74': 'home_wash_when/Before eating',
'_v75': 'home_wash_when/Before touching or preparing food',
'_v76': 'home_wash_when/Before giving food to someone else',
'_v77': 'home_wash_when/When I have dirt on my hands',
'_v78': 'home_wash_when/After touching something dirty',
'_v79': 'home_wash_when/After touching an animal',
'_v80': 'home_wash_when/After using a latrine',
'_v81': 'home_wash_when/Other',
'home_wash_when_other': 'Handwashing at home other',
'home_wash_type': 'What do you typically use to wash your hands at home?',
'_v82': 'home_wash_type/Water',
'_v83': 'home_wash_type/Ashes',
'_v84': 'home_wash_type/Sand',
'_v85': 'home_wash_type/Soap',
'_v86': 'home_wash_type/Other',
'home_wash_type_other': 'Other handwashing material at home',
'instanceID': 'instanceID',
'instanceName': 'instanceName',
'_v87': '_id',
'_v88': '_uuid',
'_v89': '_submission_time'}

```

```
[74]: df_etudiant
```

```

[74]:
      start      today      team_lead \
0  2024-05-02T12:47:28.512Z  2.030227e+12  Bati Benani
1  2024-05-02T13:24:15.292Z  2.030227e+12  Bati Benani
2  2024-05-02T12:59:28.985Z  2.030227e+12  Bati Benani
3  2024-05-02T13:38:53.439Z  2.030227e+12  Bati Benani

```

4	2024-05-02T12:51:31.353Z	2.030227e+12	Bati Benani
...
1747	2024-04-25T13:41:50.486Z	2.029622e+12	Sidina Med Mahmoud Tabou
1748	2024-04-25T13:35:51.375Z	2.029622e+12	Sidina Med Mahmoud Tabou
1749	2024-04-25T13:30:19.096Z	2.029622e+12	Sidina Med Mahmoud Tabou
1750	2024-04-25T13:17:42.627Z	2.029622e+12	Sidina Med Mahmoud Tabou
1751	2024-04-25T13:05:02.521Z	2.029622e+12	Sidina Med Mahmoud Tabou

	enum_name	gps	_v1 \
0	Ghoulam El Hadj Oumar	17.3502749 -13.8821382 84.6 3.9	17.350275
1	Cheikh Malainine Benani	17.3504233 -13.8820275 76.5 4.95	17.350423
2	Cheikh Malainine Benani	17.3504324 -13.882002 68.5 4.616	17.350432
3	Cheikh Malainine Benani	17.3504686 -13.881961 94.5 4.84	17.350469
4	Cheikh Malainine Benani	17.350499 -13.8820778 70.1 4.9	17.350499
...
1747	Oumar Anne	16.1579461 -12.7227559 61.6 5.0	16.157946
1748	Oumar Anne	16.1579475 -12.7227141 55.3 4.86	16.157947
1749	Oumar Anne	16.1579267 -12.7227983 38.8 4.966	16.157927
1750	Oumar Anne	16.1579348 -12.722745 75.5 4.933	16.157935
1751	Oumar Anne	16.158075 -12.7226967 43.6 4.983	16.158075

	_v2	_v3	_v4	region	...	_v83	_v84	_v85	_v86 \
0	-13.882138	84.6	3.900	Brakna	...	0.0	0.0	1.0	0.0
1	-13.882027	76.5	4.950	Brakna	...	0.0	0.0	1.0	0.0
2	-13.882002	68.5	4.616	Brakna	...	NaN	NaN	NaN	NaN
3	-13.881961	94.5	4.840	Brakna	...	NaN	NaN	NaN	NaN
4	-13.882078	70.1	4.900	Brakna	...	NaN	NaN	NaN	NaN
...
1747	-12.722756	61.6	5.000	Gorgol	...	0.0	0.0	0.0	0.0
1748	-12.722714	55.3	4.860	Gorgol	...	0.0	0.0	1.0	0.0
1749	-12.722798	38.8	4.966	Gorgol	...	0.0	0.0	1.0	0.0
1750	-12.722745	75.5	4.933	Gorgol	...	0.0	0.0	1.0	0.0
1751	-12.722697	43.6	4.983	Gorgol	...	0.0	0.0	1.0	0.0

	home_wash_type_other	instanceID \
0		uuid:84446fffd-96db-4ab2-afae-53884c3d0ac7
1		uuid:c9b6fa63-909a-431c-b7ed-f3241c56ebf4
2		uuid:c8685d13-5244-4aa8-813d-137ac76bdf9b
3		uuid:58aae6f6-0e87-4cb0-bc1d-e11972709d10
4		uuid:2f386798-8327-4c78-bc2c-4185e39f45eb
...
1747		uuid:fae5cba9-c8d4-42d3-a4cb-fa9b49d694ee
1748		uuid:03296f52-c19c-4b29-87b5-86a6e2007645
1749		uuid:8d744d05-d37a-415a-be2f-65b926776f99
1750		uuid:8708c4b8-ae0f-43b0-9fa4-9f71c5ef3604
1751		uuid:808ee3d5-75c9-449d-ae2c-f12b852dcc4a

		instanceName	_v87	_v88	\
0	1 - 2024-05-02	136299456	84446ffd-96db-4ab2-afae-53884c3d0ac7		
1	1 - 2024-05-02	136299330	c9b6fa63-909a-431c-b7ed-f3241c56ebf4		
2	1 - 2024-05-02	136299326	c8685d13-5244-4aa8-813d-137ac76bdf9b		
3	1 - 2024-05-02	136299331	58aae6f6-0e87-4cb0-bc1d-e11972709d10		
4	1 - 2024-05-02	136299324	2f386798-8327-4c78-bc2c-4185e39f45eb		
...		
1747	146 - 2024-04-25	135909653	fae5cba9-c8d4-42d3-a4cb-fa9b49d694ee		
1748	146 - 2024-04-25	135909650	03296f52-c19c-4b29-87b5-86a6e2007645		
1749	146 - 2024-04-25	135909645	8d744d05-d37a-415a-be2f-65b926776f99		
1750	146 - 2024-04-25	135909642	8708c4b8-ae0f-43b0-9fa4-9f71c5ef3604		
1751	146 - 2024-04-25	135909638	808ee3d5-75c9-449d-ae2c-f12b852dcc4a		

	_v89
0	2.030281e+12
1	2.030281e+12
2	2.030281e+12
3	2.030281e+12
4	2.030281e+12
...	...
1747	2.029680e+12
1748	2.029680e+12
1749	2.029680e+12
1750	2.029680e+12
1751	2.029680e+12

[1752 rows x 220 columns]

Présentation des variables de director

```
[20]: df_3 = pd.read_stata('/content/drive/MyDrive/DATA CHALLENGE/DATA CHALLENGE/
↳ Quant data before appending/director survey.dta')

df_3, meta_3 = pyreadstat.read_dta('/content/drive/MyDrive/DATA CHALLENGE/DATA_
↳ CHALLENGE/Quant data before appending/director survey.dta',
↳ encoding='latin1')
meta_3.column_names_to_labels
```

```
[20]: {'start': 'Start time',
'today': 'Today',
'team_lead': 'Name of the team leader',
'enum_name': 'Name of enumerator',
'gps': 'GPS',
'_v1': '_gps_latitude',
'_v2': '_gps_longitude',
'_v3': '_gps_altitude',
'_v4': '_gps_precision',
```

'region': 'Select the region',
 'department': 'Select the department',
 'commune': 'Select the commune',
 'school': 'Select the school',
 'consent': 'Hello! How are you? I am speaking with you today on behalf of the Future is Ours',
 'gender': 'Sex of the Director',
 'canteen_pr': 'Does your school currently have a canteen program?',
 'canteen_days': 'How many days a week does the canteen program typically provide meals?',
 'total_cooks': 'Currently, how many cooks are there in your canteen?',
 'cooks_trained': 'How many of these cooks have been trained on food preparation practices?',
 'cooks_implement': 'How many of those trained implemented these food preparation practices?',
 'management': 'Currently, including you, how many people are involved in the management of this',
 'management_train': 'How many of these people have been trained in new food storage practices?',
 'management_implement': 'Over the past year, what percentage of those people trained in the food storage',
 'cook_equipment': 'Did your school receive cooking equipment as part of the project?',
 'cook_equipment_what': 'What type of cooking equipment did the school receive?',
 '_v5': 'cook_equipment_what/Stove',
 '_v6': 'cook_equipment_what/Gas',
 '_v7': 'cook_equipment_what/Other',
 'cook_equipment_what_oth': 'If other, please specify',
 'cook_equipment_use': 'Does your school use it?',
 'cook_equipment_use_no': 'If not, what did you use instead?',
 'cook_affect': 'How did this affect food preparation?',
 '_v8': 'cook_affect/Reduces coal consumption',
 '_v9': 'cook_affect/Gets hotter than previous heat source',
 '_v10': 'cook_affect/Saves time',
 '_v11': 'cook_affect/Can prepare more food',
 '_v12': 'cook_affect/Did not affect food preparation at all',
 '_v13': 'cook_affect/Other',
 'cook_affect_other': 'If other, please specify',
 'cook_motivation': 'To what extent do you agree with the following statement: Did the project support',
 'director_motivation': 'To what extent do you agree with the following statement: The project support im',
 'med_receive': 'Did the students at your school receive anti-parasitic medicine during the 2023-',
 'book_club': 'Are there book clubs at this school?',
 'internet_access': 'Do teachers have access to internet as a part of this

project?',
 'electronic_content': 'Do teacher integrate electronic content into their teaching/lesson plans?',
 'solar_kit': 'Did your school receive a solar kit as a part of this project?',
 'coges_exist': 'Does the school have a school management committee (COGES)?',
 'coges_members': 'How many members are there on the committee?',
 'coges_women': 'How many committee members are women?',
 'coges_head': 'Is the head of the committee a man or a woman?',
 'coges_met': 'How many times have they met over the past 12 months?',
 'coges_active': 'To what extent has the committee been active over the past 12 months?',
 'covid_impact': 'Has COVID had an impact on the operation of your school in the past 12 months?',
 'covid_closure': 'Have there been extended school closures at this school during the 2023-2024 sch',
 'covid_closure_reason': 'What were the reasons?',
 'covid_days_lost': 'How many school days do you think were lost due to unplanned closures?',
 'program_satisfied': 'Overall, to what extent are you satisfied with the Future is Ours project?',
 'program_dissatisfied': 'Why are you dissatisfied?',
 'program_received': 'Which of the following did your school receive as a part of the project?',
 '_v14': 'program_received/Providing food for school meals',
 '_v15': 'program_received/Deworming medicine',
 '_v16': 'program_received/Training on food preparation, storage, sanitation, and hygiene',
 '_v17': 'program_received/Provision of cooking equipment',
 '_v18': 'program_received/Parent training on school participation',
 '_v19': 'program_received/New Arabic reading materials',
 '_v20': 'program_received/Teacher training',
 '_v21': 'program_received/Reading clubs',
 '_v22': 'program_received/Other',
 'program_received_other': 'If other, please specify',
 'enrol_data': 'Do you have enrollment data for your school from the 2023-2024 school year?',
 'data_collect': 'How do you collect enrollment data?',
 'data_quality': 'How accurate is the enrollment data?',
 'data_not_accurate': 'Why is it somewhat or not accurate?',
 'food_march_g': 'How many girls were registered in total in March?',
 'food_march_b': 'How many boys were registered in total in March?',
 'food_march': 'How many school days were there in March?',
 'food_march_g_pres': 'How many days were the girls present in total in March?',
 'food_march_b_pres': 'How many days were the boys present in total in March?',
 'food_feb_g': 'How many girls were registered in total in February?',
 'food_feb_b': 'How many boys were registered in total in February?',
 'food_feb': 'How many school days were there in February?',

'food_feb_g_pres': 'How many days were the girls present in total in February?',
 'food_feb_b_pres': 'How many days were the boys present in total in February?',
 'food_jan_g': 'How many girls were registered in total in January?',
 'food_jan_b': 'How many boys were registered in total in January?',
 'food_jan': 'How many school days were there in January?',
 'food_jan_g_pres': 'How many days were the girls present in total in January?',
 'food_jan_b_pres': 'How many days were the boys present in total in January?',
 'absence_data': 'Do you have data regarding student absences and attendance for your school from',
 'photo_register': 'Take a photo of the register',
 'girls_grade1_march': 'How many girls were registered in grade 1 in March?',
 'boys_grade1_march': 'How many boys were registered in grade 1 in March?',
 'school_march': 'How many school days were there in March?',
 'girls_grade1_abs_march': 'How many days were the girls from grade 1 absent in March?',
 'boys_grade1_abs_march': 'How many days were the boys from grade 1 absent in March?',
 'girls_grade1_pres_march': 'How many days were the girls from grade 1 present in March?',
 'boys_grade1_pres_march': 'How many days were the boys from grade 1 present in March?',
 'total_march_gr1': 'Total students in March grade1',
 'total_abs_march_gr1': 'Total absences in March grade1',
 'total_atten_march_gr1': 'Total attendance in March grade1',
 'av_atten_march_gr1': 'Average attendance in March grade1',
 'attend_march': 'What was the average rate of attendance in March?',
 'girls_grade1_feb': 'How many girls were registered in grade 1 in February?',
 'boys_grade1_feb': 'How many boys were registered in grade 1 in February?',
 'school_feb': 'How many school days were there in February?',
 'girls_grade1_abs_feb': 'How many days were the girls from grade 1 absent in February?',
 'boys_grade1_abs_feb': 'How many days were the boys from grade 1 absent in February?',
 'girls_grade1_pres_feb': 'How many days were the girls from grade 1 present in February?',
 'boys_grade1_pres_feb': 'How many days were the boys from grade 1 present in February?',
 'total_feb_gr1': 'Total students in February grade1',
 'total_abs_feb_gr1': 'Total absences in February grade1',
 'total_atten_feb_gr1': 'Total attendance in February grade1',
 'av_atten_feb_gr1': 'Average attendance in February grade1',
 'attend_feb': 'What was the average rate of attendance in February?',
 'girls_grade1_jan': 'How many girls were registered in grade 1 in January?',
 'boys_grade1_jan': 'How many boys were registered in grade 1 in January?',
 'school_jan': 'How many school days were there in January?',
 'girls_grade1_abs_jan': 'How many days were the girls from grade 1 absent in

January?',
 'boys_grade1_abs_jan': 'How many days were the boys from grade 1 absent in January?',
 'girls_grade1_pres_jan': 'How many days were the girls from grade 1 present in January?',
 'boys_grade1_pres_jan': 'How many days were the boys from grade 1 present in January?',
 'total_jan_gr1': 'Total students in January grade1',
 'total_abs_jan_gr1': 'Total absences in January grade1',
 'total_atten_jan_gr1': 'Total attendance in January grade1',
 'av_atten_jan_gr1': 'Average attendance in January grade1',
 'attend_jan': 'What was the average rate of attendance in January?',
 'girls_grade2_march': 'How many girls were registered in grade 2 in March?',
 'boys_grade2_march': 'How many boys were registered in grade 2 in March?',
 'school_march_grade2': 'How many school days were there in March?',
 'girls_grade2_abs_march': 'How many days were the girls from grade 2 absent in March?',
 'boys_grade2_abs_march': 'How many days were the boys from grade 2 absent in March?',
 'girls_grade2_pres_march': 'How many days were the girls from grade 2 present in March?',
 'boys_grade2_pres_march': 'How many days were the boys from grade 2 present in March?',
 'total_march_gr2': 'Total students in March grade2',
 'total_abs_march_gr2': 'Total absences in March grade2',
 'total_atten_march_gr2': 'Total attendance in March grade2',
 'av_atten_march_gr2': 'Average attendance in March grade2',
 'attend_march_grade2': 'What was the average rate of attendance in March?',
 'girls_grade2_feb': 'How many girls were registered in grade 2 in February?',
 'boys_grade2_feb': 'How many boys were registered in grade 2 in February?',
 'school_feb_grade2': 'How many school days were there in February?',
 'girls_grade2_abs_feb': 'How many days were the girls from grade 2 absent in February?',
 'boys_grade2_abs_feb': 'How many days were the boys from grade 2 absent in February?',
 'girls_grade2_pres_feb': 'How many days were the girls from grade 2 present in February?',
 'boys_grade2_pres_feb': 'How many days were the boys from grade 2 present in February?',
 'total_feb_gr2': 'Total students in February grade2',
 'total_abs_feb_gr2': 'Total absences in February grade2',
 'total_atten_feb_gr2': 'Total attendance in February grade2',
 'av_atten_feb_gr2': 'Average attendance in February grade2',
 'attend_feb_grade2': 'What was the average rate of attendance in February?',
 'girls_grade2_jan': 'How many girls were registered in grade 2 in January?',
 'boys_grade2_jan': 'How many boys were registered in grade 2 in January?',
 'school_jan_grade2': 'How many school days were there in January?',

'girls_grade2_abs_jan': 'How many days were the girls from grade 2 absent in January?',
 'boys_grade2_abs_jan': 'How many days were the boys from grade 2 absent in January?',
 'girls_grade2_pres_jan': 'How many days were the girls from grade 2 present in January?',
 'boys_grade2_pres_jan': 'How many days were the boys from grade 2 present in January?',
 'total_jan_gr2': 'Total students in January grade2',
 'total_abs_jan_gr2': 'Total absences in January grade2',
 'total_atten_jan_gr2': 'Total attendance in January grade2',
 'av_atten_jan_gr2': 'Average attendance in January grade2',
 'attend_jan_grade2': 'What was the average rate of attendance in January?',
 'girls_grade3_march': 'How many girls were registered in grade 3 in March?',
 'boys_grade3_march': 'How many boys were registered in grade 3 in March?',
 'school_march_grade3': 'How many school days were there in March?',
 'girls_grade3_abs_march': 'How many days were the girls from grade 3 absent in March?',
 'boys_grade3_abs_march': 'How many days were the boys from grade 3 absent in March?',
 'girls_grade3_pres_march': 'How many days were the girls from grade 3 present in March?',
 'boys_grade3_pres_march': 'How many days were the boys from grade 3 present in March?',
 'total_march_gr3': 'Total students in March grade3',
 'total_abs_march_gr3': 'Total absences in March grade3',
 'total_atten_march_gr3': 'Total attendance in March grade3',
 'av_atten_march_gr3': 'Average attendance in March grade3',
 'attend_march_grade3': 'What was the average rate of attendance in March?',
 'girls_grade3_feb': 'How many girls were registered in grade 3 in February?',
 'boys_grade3_feb': 'How many boys were registered in grade 3 in February?',
 'school_feb_grade3': 'How many school days were there in February?',
 'girls_grade3_abs_feb': 'How many days were the girls from grade 3 absent in February?',
 'boys_grade3_abs_feb': 'How many days were the boys from grade 3 absent in February?',
 'girls_grade3_pres_feb': 'How many days were the girls from grade 3 present in February?',
 'boys_grade3_pres_feb': 'How many days were the boys from grade 3 present in February?',
 'total_feb_gr3': 'Total students in February grade3',
 'total_abs_feb_gr3': 'Total absences in February grade3',
 'total_atten_feb_gr3': 'Total attendance in February grade3',
 'av_atten_feb_gr3': 'Average attendance in February grade3',
 'attend_feb_grade3': 'What was the average rate of attendance in February?',
 'girls_grade3_jan': 'How many girls were registered in grade 3 in January?',
 'boys_grade3_jan': 'How many boys were registered in grade 3 in January?',

'school_jan_grade3': 'How many school days were there in January?',
 'girls_grade3_abs_jan': 'How many days were the girls from grade 3 absent in January?',
 'boys_grade3_abs_jan': 'How many days were the boys from grade 3 absent in January?',
 'girls_grade3_pres_jan': 'How many days were the girls from grade 3 present in January?',
 'boys_grade3_pres_jan': 'How many days were the boys from grade 3 present in January?',
 'total_jan_gr3': 'Total students in January grade3',
 'total_abs_jan_gr3': 'Total absences in January grade3',
 'total_atten_jan_gr3': 'Total attendance in January grade3',
 'av_atten_jan_gr3': 'Average attendance in January grade3',
 'attend_jan_grade3': 'What was the average rate of attendance in January?',
 'girls_grade4_march': 'How many girls were registered in grade 4 in March?',
 'boys_grade4_march': 'How many boys were registered in grade 4 in March?',
 'school_march_grade4': 'How many school days were there in March?',
 'girls_grade4_abs_march': 'How many days were the girls from grade 4 absent in March?',
 'boys_grade4_abs_march': 'How many days were the boys from grade 4 absent in March?',
 'girls_grade4_pres_march': 'How many days were the girls from grade 4 present in March?',
 'boys_grade4_pres_march': 'How many days were the boys from grade 4 present in March?',
 'total_march_gr4': 'Total students in March grade4',
 'total_abs_march_gr4': 'Total absences in March grade4',
 'total_atten_march_gr4': 'Total attendance in March grade4',
 'av_atten_march_gr4': 'Average attendance in March grade4',
 'attend_march_grade4': 'What was the average rate of attendance in March?',
 'girls_grade4_feb': 'How many girls were registered in grade 4 in February?',
 'boys_grade4_feb': 'How many boys were registered in grade 4 in February?',
 'school_feb_grade4': 'How many school days were there in February?',
 'girls_grade4_abs_feb': 'How many days were the girls from grade 4 absent in February?',
 'boys_grade4_abs_feb': 'How many days were the boys from grade 4 absent in February?',
 'girls_grade4_pres_feb': 'How many days were the girls from grade 4 present in February?',
 'boys_grade4_pres_feb': 'How many days were the boys from grade 4 present in February?',
 'total_feb_gr4': 'Total students in February grade4',
 'total_abs_feb_gr4': 'Total absences in February grade4',
 'total_atten_feb_gr4': 'Total attendance in February grade4',
 'av_atten_feb_gr4': 'Average attendance in February grade4',
 'attend_feb_grade4': 'What was the average rate of attendance in February?',
 'girls_grade4_jan': 'How many girls were registered in grade 4 in January?',

'boys_grade4_jan': 'How many boys were registered in grade 4 in January?',
 'school_jan_grade4': 'How many school days were there in January?',
 'girls_grade4_abs_jan': 'How many days were the girls from grade 4 absent in January?',
 'boys_grade4_abs_jan': 'How many days were the boys from grade 4 absent in January?',
 'girls_grade4_pres_jan': 'How many days were the girls from grade 4 present in January?',
 'boys_grade4_pres_jan': 'How many days were the boys from grade 4 present in January?',
 'total_jan_gr4': 'Total students in January grade4',
 'total_abs_jan_gr4': 'Total absences in January grade4',
 'total_atten_jan_gr4': 'Total attendance in January grade4',
 'av_atten_jan_gr4': 'Average attendance in January grade4',
 'attend_jan_grade4': 'What was the average rate of attendance in January?',
 'girls_grade5_march': 'How many girls were registered in grade 5 in March?',
 'boys_grade5_march': 'How many boys were registered in grade 5 in March?',
 'school_march_grade5': 'How many school days were there in March?',
 'girls_grade5_abs_march': 'How many days were the girls from grade 5 absent in March?',
 'boys_grade5_abs_march': 'How many days were the boys from grade 5 absent in March?',
 'girls_grade5_pres_march': 'How many days were the girls from grade 5 present in March?',
 'boys_grade5_pres_march': 'How many days were the boys from grade 5 present in March?',
 'total_march_gr5': 'Total students in March grade5',
 'total_abs_march_gr5': 'Total absences in March grade5',
 'total_atten_march_gr5': 'Total attendance in March grade5',
 'av_atten_march_gr5': 'Average attendance in March grade5',
 'attend_march_grade5': 'What was the average rate of attendance in March?',
 'girls_grade5_feb': 'How many girls were registered in grade 5 in February?',
 'boys_grade5_feb': 'How many boys were registered in grade 5 in February?',
 'school_feb_grade5': 'How many school days were there in February?',
 'girls_grade5_abs_feb': 'How many days were the girls from grade 5 absent in February?',
 'boys_grade5_abs_feb': 'How many days were the boys from grade 5 absent in February?',
 'girls_grade5_pres_feb': 'How many days were the girls from grade 5 present in February?',
 'boys_grade5_pres_feb': 'How many days were the boys from grade 5 present in February?',
 'total_feb_gr5': 'Total students in February grade5',
 'total_abs_feb_gr5': 'Total absences in February grade5',
 'total_atten_feb_gr5': 'Total attendance in February grade5',
 'av_atten_feb_gr5': 'Average attendance in February grade5',
 'attend_feb_grade5': 'What was the average rate of attendance in February?',

'girls_grade5_jan': 'How many girls were registered in grade 5 in January?',
 'boys_grade5_jan': 'How many boys were registered in grade 5 in January?',
 'school_jan_grade5': 'How many school days were there in January?',
 'girls_grade5_abs_jan': 'How many days were the girls from grade 5 absent in January?',
 'boys_grade5_abs_jan': 'How many days were the boys from grade 5 absent in January?',
 'girls_grade5_pres_jan': 'How many days were the girls from grade 5 present in January?',
 'boys_grade5_pres_jan': 'How many days were the boys from grade 5 present in January?',
 'total_jan_gr5': 'Total students in January grade5',
 'total_abs_jan_gr5': 'Total absences in January grade5',
 'total_atten_jan_gr5': 'Total attendance in January grade5',
 'av_atten_jan_gr5': 'Average attendance in January grade5',
 'attend_jan_grade5': 'What was the average rate of attendance in January?',
 'girls_grade6_march': 'How many girls were registered in grade 6 in March?',
 'boys_grade6_march': 'How many boys were registered in grade 6 in March?',
 'school_march_grade6': 'How many school days were there in March?',
 'girls_grade6_abs_march': 'How many days were the girls from grade 6 absent in March?',
 'boys_grade6_abs_march': 'How many days were the boys from grade 6 absent in March?',
 'girls_grade6_pres_march': 'How many days were the girls from grade 6 present in March?',
 'boys_grade6_pres_march': 'How many days were the boys from grade 6 present in March?',
 'total_march_gr6': 'Total students in March grade6',
 'total_abs_march_gr6': 'Total absences in March grade6',
 'total_atten_march_gr6': 'Total attendance in March grade6',
 'av_atten_march_gr6': 'Average attendance in March grade6',
 'attend_march_grade6': 'What was the average rate of attendance in March?',
 'girls_grade6_feb': 'How many girls were registered in grade 6 in February?',
 'boys_grade6_feb': 'How many boys were registered in grade 6 in February?',
 'school_feb_grade6': 'How many school days were there in February?',
 'girls_grade6_abs_feb': 'How many days were the girls from grade 6 absent in February?',
 'boys_grade6_abs_feb': 'How many days were the boys from grade 6 absent in February?',
 'girls_grade6_pres_feb': 'How many days were the girls from grade 6 present in February?',
 'boys_grade6_pres_feb': 'How many days were the boys from grade 6 present in February?',
 'total_feb_gr6': 'Total students in February grade6',
 'total_abs_feb_gr6': 'Total absences in February grade6',
 'total_atten_feb_gr6': 'Total attendance in February grade6',
 'av_atten_feb_gr6': 'Average attendance in February grade6',

'attend_feb_grade6': 'What was the average rate of attendance in February?',
 'girls_grade6_jan': 'How many girls were registered in grade 6 in January?',
 'boys_grade6_jan': 'How many boys were registered in grade 6 in January?',
 'school_jan_grade6': 'How many school days were there in January?',
 'girls_grade6_abs_jan': 'How many days were the girls from grade 6 absent in January?',
 'boys_grade6_abs_jan': 'How many days were the boys from grade 6 absent in January?',
 'girls_grade6_pres_jan': 'How many days were the girls from grade 6 present in January?',
 'boys_grade6_pres_jan': 'How many days were the boys from grade 6 present in January?',
 'total_jan_gr6': 'Total students in January grade6',
 'total_abs_jan_gr6': 'Total absences in January grade6',
 'total_atten_jan_gr6': 'Total attendance in January grade6',
 'av_atten_jan_gr6': 'Average attendance in January grade6',
 'attend_jan_grade56': 'What was the average rate of attendance in January?',
 'at_data_collect': 'How do you collect attendance data?',
 'at_data_quality': 'How accurate is the attendance data you have provided?',
 'at_data_not_accurate': 'Why is it somewhat or not accurate?',
 'total_latrines': 'How many latrines does the school have for students?',
 'latrines_separate': 'Are there separately marked latrines for boys and girls?',
 'latrine_clean': 'How clean are the latrines?',
 'latrine_maintain': 'How well maintained is the latrine structure?',
 'latrine_water': 'Are the latrines connected to water?',
 'latrine_water_work': 'Does the water connection work?',
 'hand_wash_source': 'What is the water source used for hand washing in the latrines?',
 'hand_wash_source_oth': 'If other, please specify',
 'photo_handwash': 'Take a picture of the handwashing water source',
 'water_source_clean': 'To what extent is the water source clean?',
 'latrine_soap': 'Is there soap?',
 'latrine_towel': 'Is there a clean towel or other material for drying hands?',
 'latrine_private': 'Does the latrine preserve privacy (can the door be completely closed so that peo',
 'latrine_boys': 'Now that you have inspected the girls' latrine, check the boy's latrine. Are the',
 'main_drink_source': 'What is the main source of water for drinking at the school?',
 'main_drink_source_oth': 'If other, please specify',
 'photo_drink_source': 'Take a picture of the water source for the school',
 'water_issues': 'Try and get water as if you were student trying to get something to drink. Do yo',
 'water_issues_what': 'What are the problems with the water?',
 '_v23': 'water_issues_what/Limited quantity',
 '_v24': 'water_issues_what/Water takes time to come',


```

'_v25': 'water_issues_what/Equipment is old/not working',
'_v26': 'water_issues_what/Water does not look clean',
'_v27': 'water_issues_what/Area around the water is not clean',
'_v28': 'water_issues_what/Other',
'water_issues_other': 'If other, please specify',
'water_on_property': 'Is the water source on school property?',
'water_walk': 'How many minutes walk is the water source from the school?',
'add_info': 'FOR ENUMERATOR: Please note any additional information that is
important for us',
'instanceID': 'instanceID',
'instanceName': 'instanceName',
'_v29': '_id',
'_v30': '_uuid',
'_v31': '_submission_time'}

```

[21]: df_3

```

[21]:
      start      today  team_lead  enum_name  \
0  2024-05-02T13:00:33.685Z  2.030227e+12      3      301
1  2024-04-22T10:27:49.144Z  2.029363e+12      5      501
2  2024-04-24T13:46:03.475Z  2.029536e+12      5      501
3  2024-04-24T11:27:07.968Z  2.029536e+12      5      501
4  2024-04-23T08:27:12.662Z  2.029450e+12      5      501
..      ...      ...      ...      ...
141 2024-04-26T08:17:30.031Z  2.029709e+12      4      401
142 2024-04-26T11:57:33.149Z  2.029709e+12      4      401
143 2024-04-29T08:28:06.303Z  2.029968e+12      4      401
144 2024-04-26T10:29:39.883Z  2.029709e+12      4      401
145 2024-04-25T13:22:11.645Z  2.029622e+12      4      401

      gps      _v1      _v2      _v3      _v4  \
0  17.3505286 -13.8823365 75.3 4.5 17.350529 -13.882336 75.3 4.500
1  17.1219157 -14.0669924 76.0 4.88 17.121916 -14.066992 76.0 4.880
2  17.0472362 -13.8998554 75.2 4.866 17.047236 -13.899855 75.2 4.866
3  17.0289048 -13.9666087 74.1 4.7 17.028905 -13.966609 74.1 4.700
4  17.097394 -13.8910092 86.2 4.671 17.097394 -13.891009 86.2 4.671
..      ...      ...      ...      ...      ...
141 16.3740633 -13.1227434 54.6 5.0 16.374063 -13.122743 54.6 5.000
142 16.4119209 -13.135757 90.3 4.883 16.411921 -13.135757 90.3 4.883
143 16.4185709 -13.1366271 73.0 3.666 16.418571 -13.136627 73.0 3.666
144 16.4066204 -13.1362544 80.5 4.78 16.406620 -13.136254 80.5 4.780
145 16.158213 -12.7228823 62.5 4.966 16.158213 -12.722882 62.5 4.966

      region  ...  _v28      water_issues_other  water_on_property  \
0          1  ...   NaN                        1
1          1  ...     0                        1
2          1  ...     0                        1

```

3	1	...	NaN		1
4	1	...	1	Il n'ya pas de problÃme d'eau	1
..
141	2	...	NaN		1
142	2	...	0		1
143	2	...	NaN		1
144	2	...	NaN		1
145	2	...	NaN		1

	water_walk		add_info	\
0	1	- pas de rÃ©fectoire \n- clÃture en mauvais Ã...		
1	1	Il y a un fut Ã eau oÃ les ÃlÃves boivent		
2	1	Il ya des lavabo qui ne sont plus utilisÃs c...		
3	1	Les latrines sont rarement utilisÃes par les ...		
4	1	L'Ãcole a le problÃme de rÃ©fectoire oÃ les...		
..		
141	1	Pas d'informations supplÃ©mentaires		
142	1	Pas d'informations supplÃ©mentaires		
143	1	Pas d'observation		
144	1	Manque de magasin pour la conservation des ali...		
145	1	Une fois sur place Ã sabahalla, on a constatÃ...		

	instanceID	instanceName	_v29	\
0	uuid:ab6c7b35-b8fd-4ba8-bb3f-30a3479bca76	1 - 2024-05-02	136299334	
1	uuid:d388ad9f-226a-4851-b4fe-ef97203fdee5	2 - 2024-04-22	135786482	
2	uuid:80cf7f21-0ca6-4bd5-bb32-b10bdbcfb033	3 - 2024-04-24	135940688	
3	uuid:b1dc2d4c-6798-4bac-8146-0c4c21692da0	4 - 2024-04-24	135940697	
4	uuid:1efdc50a-8d84-411d-99d3-453255f9156b	5 - 2024-04-23	135940701	
..	
141	uuid:87bbb303-0386-4fb6-a5d1-890b0ae3f329	142 - 2024-04-26	135943890	
142	uuid:fd493968-0c38-4f11-b441-86f53a150dc4	143 - 2024-04-26	135943941	
143	uuid:2d906482-caa7-47d1-9821-3a393e56a833	144 - 2024-04-29	136096825	
144	uuid:34ab1cad-bea6-4851-bd84-787d5742529c	145 - 2024-04-26	135943951	
145	uuid:35ac61d5-17a0-4209-b23b-068d90daf01d	146 - 2024-04-25	135916800	

	_v30	_v31
0	ab6c7b35-b8fd-4ba8-bb3f-30a3479bca76	2.030281e+12
1	d388ad9f-226a-4851-b4fe-ef97203fdee5	2.029477e+12
2	80cf7f21-0ca6-4bd5-bb32-b10bdbcfb033	2.029763e+12
3	b1dc2d4c-6798-4bac-8146-0c4c21692da0	2.029763e+12
4	1efdc50a-8d84-411d-99d3-453255f9156b	2.029763e+12
..
141	87bbb303-0386-4fb6-a5d1-890b0ae3f329	2.029769e+12
142	fd493968-0c38-4f11-b441-86f53a150dc4	2.029769e+12
143	2d906482-caa7-47d1-9821-3a393e56a833	2.030007e+12
144	34ab1cad-bea6-4851-bd84-787d5742529c	2.029769e+12
145	35ac61d5-17a0-4209-b23b-068d90daf01d	2.029694e+12

[146 rows x 343 columns]

```
[22]: df_4 = pd.read_stata('/content/drive/MyDrive/DATA CHALLENGE/DATA CHALLENGE/
↳Quant data before appending/director survey.dta')

df_4, meta_4 = pyreadstat.read_dta('/content/drive/MyDrive/DATA CHALLENGE/DATA_
↳CHALLENGE/Quant data before appending/director survey.dta',
↳encoding='latin1')
meta_4.column_names_to_labels
```

```
[22]: {'start': 'Start time',
      'today': 'Today',
      'team_lead': 'Name of the team leader',
      'enum_name': 'Name of enumerator',
      'gps': 'GPS',
      '_v1': '_gps_latitude',
      '_v2': '_gps_longitude',
      '_v3': '_gps_altitude',
      '_v4': '_gps_precision',
      'region': 'Select the region',
      'department': 'Select the department',
      'commune': 'Select the commune',
      'school': 'Select the school',
      'consent': 'Hello! How are you? I am speaking with you today on behalf of the
Future is Ours',
      'gender': 'Sex of the Director',
      'canteen_pr': 'Does your school currently have a canteen program?',
      'canteen_days': 'How many days a week does the canteen program typically
provide meals?',
      'total_cooks': 'Currently, how many cooks are there in your canteen?',
      'cooks_trained': 'How many of these cooks have been trained on food preparation
practices?',
      'cooks_implement': 'How many of those trained implemented these food
preparation practices?',
      'management': 'Currently, including you, how many people are involved in the
management of this',
      'management_train': 'How many of these people have been trained in new food
storage practices?',
      'management_implement': 'Over the past year, what percentage of those people
trained in the food storage',
      'cook_equipment': 'Did your school receive cooking equipment as part of the
project?',
      'cook_equipment_what': 'What type of cooking equipment did the school
receive?',
      '_v5': 'cook_equipment_what/Stove',
      '_v6': 'cook_equipment_what/Gas',
```

'_v7': 'cook_equipment_what/Other',
 'cook_equipment_what_oth': 'If other, please specify',
 'cook_equipment_use': 'Does your school use it?',
 'cook_equipment_use_no': 'If not, what did you use instead?',
 'cook_affect': 'How did this affect food preparation?',
 '_v8': 'cook_affect/Reduces coal consumption',
 '_v9': 'cook_affect/Gets hotter than previous heat source',
 '_v10': 'cook_affect/Saves time',
 '_v11': 'cook_affect/Can prepare more food',
 '_v12': 'cook_affect/Did not affect food preparation at all',
 '_v13': 'cook_affect/Other',
 'cook_affect_other': 'If other, please specify',
 'cook_motivation': 'To what extent do you agree with the following statement:
 Did the project suppor',
 'director_motivation': 'To what extent do you agree with the following
 statement: The project support im',
 'med_receive': 'Did the students at your school receive anti-parasitic medicine
 during the 2023-',
 'book_club': 'Are there book clubs at this school?',
 'internet_access': 'Do teachers have access to internet as a part of this
 project?',
 'electronic_content': 'Do teacher integrate electronic content into their
 teaching/lesson plans?',
 'solar_kit': 'Did your school receive a solar kit as a part of this project?',
 'coges_exist': 'Does the school have a school management committee (COGES)?',
 'coges_members': 'How many members are there on the committee?',
 'coges_women': 'How many committee members are women?',
 'coges_head': 'Is the head of the committee a man or a woman?',
 'coges_met': 'How many times have they met over the past 12 months?',
 'coges_active': 'To what extent has the committee been active over the past 12
 months?',
 'covid_impact': 'Has COVID had an impact on the operation of your school in the
 past 12 months?',
 'covid_closure': 'Have there been extended school closures at this school
 during the 2023-2024 sch',
 'covid_closure_reason': 'What were the reasons?',
 'covid_days_lost': 'How many school days do you think were lost due to
 unplanned closures?',
 'program_satisfied': 'Overall, to what extent are you satisfied with the Future
 is Ours project?',
 'program_dissatisfied': 'Why are you dissatisfied?',
 'program_received': 'Which of the following did your school receive as a part
 of the project?',
 '_v14': 'program_received/Providing food for school meals',
 '_v15': 'program_received/Deworming medicine',
 '_v16': 'program_received/Training on food preparation, storage, sanitation,
 and hygiene',

'_v17': 'program_received/Provision of cooking equipment',
 '_v18': 'program_received/Parent training on school participation',
 '_v19': 'program_received/New Arabic reading materials',
 '_v20': 'program_received/Teacher training',
 '_v21': 'program_received/Reading clubs',
 '_v22': 'program_received/Other',
 'program_received_other': 'If other, please specify',
 'enrol_data': 'Do you have enrollment data for your school from the 2023-2024 school year?',
 'data_collect': 'How do you collect enrollment data?',
 'data_quality': 'How accurate is the enrollment data?',
 'data_not_accurate': 'Why is it somewhat or not accurate?',
 'food_march_g': 'How many girls were registered in total in March?',
 'food_march_b': 'How many boys were registered in total in March?',
 'food_march': 'How many school days were there in March?',
 'food_march_g_pres': 'How many days were the girls present in total in March?',
 'food_march_b_pres': 'How many days were the boys present in total in March?',
 'food_feb_g': 'How many girls were registered in total in February?',
 'food_feb_b': 'How many boys were registered in total in February?',
 'food_feb': 'How many school days were there in February?',
 'food_feb_g_pres': 'How many days were the girls present in total in February?',
 'food_feb_b_pres': 'How many days were the boys present in total in February?',
 'food_jan_g': 'How many girls were registered in total in January?',
 'food_jan_b': 'How many boys were registered in total in January?',
 'food_jan': 'How many school days were there in January?',
 'food_jan_g_pres': 'How many days were the girls present in total in January?',
 'food_jan_b_pres': 'How many days were the boys present in total in January?',
 'absence_data': 'Do you have data regarding student absences and attendance for your school from',
 'photo_register': 'Take a photo of the register',
 'girls_grade1_march': 'How many girls were registered in grade 1 in March?',
 'boys_grade1_march': 'How many boys were registered in grade 1 in March?',
 'school_march': 'How many school days were there in March?',
 'girls_grade1_abs_march': 'How many days were the girls from grade 1 absent in March?',
 'boys_grade1_abs_march': 'How many days were the boys from grade 1 absent in March?',
 'girls_grade1_pres_march': 'How many days were the girls from grade 1 present in March?',
 'boys_grade1_pres_march': 'How many days were the boys from grade 1 present in March?',
 'total_march_gr1': 'Total students in March grade1',
 'total_abs_march_gr1': 'Total absences in March grade1',
 'total_atten_march_gr1': 'Total attendance in March grade1',
 'av_atten_march_gr1': 'Average attendance in March grade1',
 'attend_march': 'What was the average rate of attendance in March?',

'girls_grade1_feb': 'How many girls were registered in grade 1 in February?',
 'boys_grade1_feb': 'How many boys were registered in grade 1 in February?',
 'school_feb': 'How many school days were there in February?',
 'girls_grade1_abs_feb': 'How many days were the girls from grade 1 absent in February?',
 'boys_grade1_abs_feb': 'How many days were the boys from grade 1 absent in February?',
 'girls_grade1_pres_feb': 'How many days were the girls from grade 1 present in February?',
 'boys_grade1_pres_feb': 'How many days were the boys from grade 1 present in February?',
 'total_feb_gr1': 'Total students in February grade1',
 'total_abs_feb_gr1': 'Total absences in February grade1',
 'total_atten_feb_gr1': 'Total attendance in February grade1',
 'av_atten_feb_gr1': 'Average attendance in February grade1',
 'attend_feb': 'What was the average rate of attendance in February?',
 'girls_grade1_jan': 'How many girls were registered in grade 1 in January?',
 'boys_grade1_jan': 'How many boys were registered in grade 1 in January?',
 'school_jan': 'How many school days were there in January?',
 'girls_grade1_abs_jan': 'How many days were the girls from grade 1 absent in January?',
 'boys_grade1_abs_jan': 'How many days were the boys from grade 1 absent in January?',
 'girls_grade1_pres_jan': 'How many days were the girls from grade 1 present in January?',
 'boys_grade1_pres_jan': 'How many days were the boys from grade 1 present in January?',
 'total_jan_gr1': 'Total students in January grade1',
 'total_abs_jan_gr1': 'Total absences in January grade1',
 'total_atten_jan_gr1': 'Total attendance in January grade1',
 'av_atten_jan_gr1': 'Average attendance in January grade1',
 'attend_jan': 'What was the average rate of attendance in January?',
 'girls_grade2_march': 'How many girls were registered in grade 2 in March?',
 'boys_grade2_march': 'How many boys were registered in grade 2 in March?',
 'school_march_grade2': 'How many school days were there in March?',
 'girls_grade2_abs_march': 'How many days were the girls from grade 2 absent in March?',
 'boys_grade2_abs_march': 'How many days were the boys from grade 2 absent in March?',
 'girls_grade2_pres_march': 'How many days were the girls from grade 2 present in March?',
 'boys_grade2_pres_march': 'How many days were the boys from grade 2 present in March?',
 'total_march_gr2': 'Total students in March grade2',
 'total_abs_march_gr2': 'Total absences in March grade2',
 'total_atten_march_gr2': 'Total attendance in March grade2',
 'av_atten_march_gr2': 'Average attendance in March grade2',

'attend_march_grade2': 'What was the average rate of attendance in March?',
 'girls_grade2_feb': 'How many girls were registered in grade 2 in February?',
 'boys_grade2_feb': 'How many boys were registered in grade 2 in February?',
 'school_feb_grade2': 'How many school days were there in February?',
 'girls_grade2_abs_feb': 'How many days were the girls from grade 2 absent in February?',
 'boys_grade2_abs_feb': 'How many days were the boys from grade 2 absent in February?',
 'girls_grade2_pres_feb': 'How many days were the girls from grade 2 present in February?',
 'boys_grade2_pres_feb': 'How many days were the boys from grade 2 present in February?',
 'total_feb_gr2': 'Total students in February grade2',
 'total_abs_feb_gr2': 'Total absences in February grade2',
 'total_atten_feb_gr2': 'Total attendance in February grade2',
 'av_atten_feb_gr2': 'Average attendance in February grade2',
 'attend_feb_grade2': 'What was the average rate of attendance in February?',
 'girls_grade2_jan': 'How many girls were registered in grade 2 in January?',
 'boys_grade2_jan': 'How many boys were registered in grade 2 in January?',
 'school_jan_grade2': 'How many school days were there in January?',
 'girls_grade2_abs_jan': 'How many days were the girls from grade 2 absent in January?',
 'boys_grade2_abs_jan': 'How many days were the boys from grade 2 absent in January?',
 'girls_grade2_pres_jan': 'How many days were the girls from grade 2 present in January?',
 'boys_grade2_pres_jan': 'How many days were the boys from grade 2 present in January?',
 'total_jan_gr2': 'Total students in January grade2',
 'total_abs_jan_gr2': 'Total absences in January grade2',
 'total_atten_jan_gr2': 'Total attendance in January grade2',
 'av_atten_jan_gr2': 'Average attendance in January grade2',
 'attend_jan_grade2': 'What was the average rate of attendance in January?',
 'girls_grade3_march': 'How many girls were registered in grade 3 in March?',
 'boys_grade3_march': 'How many boys were registered in grade 3 in March?',
 'school_march_grade3': 'How many school days were there in March?',
 'girls_grade3_abs_march': 'How many days were the girls from grade 3 absent in March?',
 'boys_grade3_abs_march': 'How many days were the boys from grade 3 absent in March?',
 'girls_grade3_pres_march': 'How many days were the girls from grade 3 present in March?',
 'boys_grade3_pres_march': 'How many days were the boys from grade 3 present in March?',
 'total_march_gr3': 'Total students in March grade3',
 'total_abs_march_gr3': 'Total absences in March grade3',
 'total_atten_march_gr3': 'Total attendance in March grade3',

'av_atten_march_gr3': 'Average attendance in March grade3',
 'attend_march_grade3': 'What was the average rate of attendance in March?',
 'girls_grade3_feb': 'How many girls were registered in grade 3 in February?',
 'boys_grade3_feb': 'How many boys were registered in grade 3 in February?',
 'school_feb_grade3': 'How many school days were there in February?',
 'girls_grade3_abs_feb': 'How many days were the girls from grade 3 absent in February?',
 'boys_grade3_abs_feb': 'How many days were the boys from grade 3 absent in February?',
 'girls_grade3_pres_feb': 'How many days were the girls from grade 3 present in February?',
 'boys_grade3_pres_feb': 'How many days were the boys from grade 3 present in February?',
 'total_feb_gr3': 'Total students in February grade3',
 'total_abs_feb_gr3': 'Total absences in February grade3',
 'total_atten_feb_gr3': 'Total attendance in February grade3',
 'av_atten_feb_gr3': 'Average attendance in February grade3',
 'attend_feb_grade3': 'What was the average rate of attendance in February?',
 'girls_grade3_jan': 'How many girls were registered in grade 3 in January?',
 'boys_grade3_jan': 'How many boys were registered in grade 3 in January?',
 'school_jan_grade3': 'How many school days were there in January?',
 'girls_grade3_abs_jan': 'How many days were the girls from grade 3 absent in January?',
 'boys_grade3_abs_jan': 'How many days were the boys from grade 3 absent in January?',
 'girls_grade3_pres_jan': 'How many days were the girls from grade 3 present in January?',
 'boys_grade3_pres_jan': 'How many days were the boys from grade 3 present in January?',
 'total_jan_gr3': 'Total students in January grade3',
 'total_abs_jan_gr3': 'Total absences in January grade3',
 'total_atten_jan_gr3': 'Total attendance in January grade3',
 'av_atten_jan_gr3': 'Average attendance in January grade3',
 'attend_jan_grade3': 'What was the average rate of attendance in January?',
 'girls_grade4_march': 'How many girls were registered in grade 4 in March?',
 'boys_grade4_march': 'How many boys were registered in grade 4 in March?',
 'school_march_grade4': 'How many school days were there in March?',
 'girls_grade4_abs_march': 'How many days were the girls from grade 4 absent in March?',
 'boys_grade4_abs_march': 'How many days were the boys from grade 4 absent in March?',
 'girls_grade4_pres_march': 'How many days were the girls from grade 4 present in March?',
 'boys_grade4_pres_march': 'How many days were the boys from grade 4 present in March?',
 'total_march_gr4': 'Total students in March grade4',
 'total_abs_march_gr4': 'Total absences in March grade4',

'total_atten_march_gr4': 'Total attendance in March grade4',
 'av_atten_march_gr4': 'Average attendance in March grade4',
 'attend_march_grade4': 'What was the average rate of attendance in March?',
 'girls_grade4_feb': 'How many girls were registered in grade 4 in February?',
 'boys_grade4_feb': 'How many boys were registered in grade 4 in February?',
 'school_feb_grade4': 'How many school days were there in February?',
 'girls_grade4_abs_feb': 'How many days were the girls from grade 4 absent in February?',
 'boys_grade4_abs_feb': 'How many days were the boys from grade 4 absent in February?',
 'girls_grade4_pres_feb': 'How many days were the girls from grade 4 present in February?',
 'boys_grade4_pres_feb': 'How many days were the boys from grade 4 present in February?',
 'total_feb_gr4': 'Total students in February grade4',
 'total_abs_feb_gr4': 'Total absences in February grade4',
 'total_atten_feb_gr4': 'Total attendance in February grade4',
 'av_atten_feb_gr4': 'Average attendance in February grade4',
 'attend_feb_grade4': 'What was the average rate of attendance in February?',
 'girls_grade4_jan': 'How many girls were registered in grade 4 in January?',
 'boys_grade4_jan': 'How many boys were registered in grade 4 in January?',
 'school_jan_grade4': 'How many school days were there in January?',
 'girls_grade4_abs_jan': 'How many days were the girls from grade 4 absent in January?',
 'boys_grade4_abs_jan': 'How many days were the boys from grade 4 absent in January?',
 'girls_grade4_pres_jan': 'How many days were the girls from grade 4 present in January?',
 'boys_grade4_pres_jan': 'How many days were the boys from grade 4 present in January?',
 'total_jan_gr4': 'Total students in January grade4',
 'total_abs_jan_gr4': 'Total absences in January grade4',
 'total_atten_jan_gr4': 'Total attendance in January grade4',
 'av_atten_jan_gr4': 'Average attendance in January grade4',
 'attend_jan_grade4': 'What was the average rate of attendance in January?',
 'girls_grade5_march': 'How many girls were registered in grade 5 in March?',
 'boys_grade5_march': 'How many boys were registered in grade 5 in March?',
 'school_march_grade5': 'How many school days were there in March?',
 'girls_grade5_abs_march': 'How many days were the girls from grade 5 absent in March?',
 'boys_grade5_abs_march': 'How many days were the boys from grade 5 absent in March?',
 'girls_grade5_pres_march': 'How many days were the girls from grade 5 present in March?',
 'boys_grade5_pres_march': 'How many days were the boys from grade 5 present in March?',
 'total_march_gr5': 'Total students in March grade5',

'total_abs_march_gr5': 'Total absences in March grade5',
 'total_atten_march_gr5': 'Total attendance in March grade5',
 'av_atten_march_gr5': 'Average attendance in March grade5',
 'attend_march_grade5': 'What was the average rate of attendance in March?',
 'girls_grade5_feb': 'How many girls were registered in grade 5 in February?',
 'boys_grade5_feb': 'How many boys were registered in grade 5 in February?',
 'school_feb_grade5': 'How many school days were there in February?',
 'girls_grade5_abs_feb': 'How many days were the girls from grade 5 absent in February?',
 'boys_grade5_abs_feb': 'How many days were the boys from grade 5 absent in February?',
 'girls_grade5_pres_feb': 'How many days were the girls from grade 5 present in February?',
 'boys_grade5_pres_feb': 'How many days were the boys from grade 5 present in February?',
 'total_feb_gr5': 'Total students in February grade5',
 'total_abs_feb_gr5': 'Total absences in February grade5',
 'total_atten_feb_gr5': 'Total attendance in February grade5',
 'av_atten_feb_gr5': 'Average attendance in February grade5',
 'attend_feb_grade5': 'What was the average rate of attendance in February?',
 'girls_grade5_jan': 'How many girls were registered in grade 5 in January?',
 'boys_grade5_jan': 'How many boys were registered in grade 5 in January?',
 'school_jan_grade5': 'How many school days were there in January?',
 'girls_grade5_abs_jan': 'How many days were the girls from grade 5 absent in January?',
 'boys_grade5_abs_jan': 'How many days were the boys from grade 5 absent in January?',
 'girls_grade5_pres_jan': 'How many days were the girls from grade 5 present in January?',
 'boys_grade5_pres_jan': 'How many days were the boys from grade 5 present in January?',
 'total_jan_gr5': 'Total students in January grade5',
 'total_abs_jan_gr5': 'Total absences in January grade5',
 'total_atten_jan_gr5': 'Total attendance in January grade5',
 'av_atten_jan_gr5': 'Average attendance in January grade5',
 'attend_jan_grade5': 'What was the average rate of attendance in January?',
 'girls_grade6_march': 'How many girls were registered in grade 6 in March?',
 'boys_grade6_march': 'How many boys were registered in grade 6 in March?',
 'school_march_grade6': 'How many school days were there in March?',
 'girls_grade6_abs_march': 'How many days were the girls from grade 6 absent in March?',
 'boys_grade6_abs_march': 'How many days were the boys from grade 6 absent in March?',
 'girls_grade6_pres_march': 'How many days were the girls from grade 6 present in March?',
 'boys_grade6_pres_march': 'How many days were the boys from grade 6 present in March?',

'total_march_gr6': 'Total students in March grade6',
 'total_abs_march_gr6': 'Total absences in March grade6',
 'total_atten_march_gr6': 'Total attendance in March grade6',
 'av_atten_march_gr6': 'Average attendance in March grade6',
 'attend_march_grade6': 'What was the average rate of attendance in March?',
 'girls_grade6_feb': 'How many girls were registered in grade 6 in February?',
 'boys_grade6_feb': 'How many boys were registered in grade 6 in February?',
 'school_feb_grade6': 'How many school days were there in February?',
 'girls_grade6_abs_feb': 'How many days were the girls from grade 6 absent in February?',
 'boys_grade6_abs_feb': 'How many days were the boys from grade 6 absent in February?',
 'girls_grade6_pres_feb': 'How many days were the girls from grade 6 present in February?',
 'boys_grade6_pres_feb': 'How many days were the boys from grade 6 present in February?',
 'total_feb_gr6': 'Total students in February grade6',
 'total_abs_feb_gr6': 'Total absences in February grade6',
 'total_atten_feb_gr6': 'Total attendance in February grade6',
 'av_atten_feb_gr6': 'Average attendance in February grade6',
 'attend_feb_grade6': 'What was the average rate of attendance in February?',
 'girls_grade6_jan': 'How many girls were registered in grade 6 in January?',
 'boys_grade6_jan': 'How many boys were registered in grade 6 in January?',
 'school_jan_grade6': 'How many school days were there in January?',
 'girls_grade6_abs_jan': 'How many days were the girls from grade 6 absent in January?',
 'boys_grade6_abs_jan': 'How many days were the boys from grade 6 absent in January?',
 'girls_grade6_pres_jan': 'How many days were the girls from grade 6 present in January?',
 'boys_grade6_pres_jan': 'How many days were the boys from grade 6 present in January?',
 'total_jan_gr6': 'Total students in January grade6',
 'total_abs_jan_gr6': 'Total absences in January grade6',
 'total_atten_jan_gr6': 'Total attendance in January grade6',
 'av_atten_jan_gr6': 'Average attendance in January grade6',
 'attend_jan_grade6': 'What was the average rate of attendance in January?',
 'at_data_collect': 'How do you collect attendance data?',
 'at_data_quality': 'How accurate is the attendance data you have provided?',
 'at_data_not_accurate': 'Why is it somewhat or not accurate?',
 'total_latrines': 'How many latrines does the school have for students?',
 'latrines_separate': 'Are there separately marked latrines for boys and girls?',
 'latrine_clean': 'How clean are the latrines?',
 'latrine_maintain': 'How well maintained is the latrine structure?',
 'latrine_water': 'Are the latrines connected to water?',
 'latrine_water_work': 'Does the water connection work?',

```

'hand_wash_source': 'What is the water source used for hand washing in the
latrines?',
'hand_wash_source_oth': 'If other, please specify',
'photo_handwash': 'Take a picture of the handwashing water source',
'water_source_clean': 'To what extent is the water source clean?',
'latrine_soap': 'Is there soap?',
'latrine_towel': 'Is there a clean towel or other material for drying hands?',
'latrine_private': 'Does the latrine preserve privacy (can the door be
completely closed so that peo',
'latrine_boys': "Now that you have inspected the girls' latrine, check the
boy's latrine. Are the",
'main_drink_source': 'What is the main source of water for drinking at the
school?',
'main_drink_source_oth': 'If other, please specify',
'photo_drink_source': 'Take a picture of the water source for the school',
'water_issues': 'Try and get water as if you were student trying to get
something to drink. Do yo',
'water_issues_what': 'What are the problems with the water?',
'_v23': 'water_issues_what/Limited quantity',
'_v24': 'water_issues_what/Water takes time to come',
'_v25': 'water_issues_what/Equipment is old/not working',
'_v26': 'water_issues_what/Water does not look clean',
'_v27': 'water_issues_what/Area around the water is not clean',
'_v28': 'water_issues_what/Other',
'water_issues_other': 'If other, please specify',
'water_on_property': 'Is the water source on school property?',
'water_walk': 'How many minutes walk is the water source from the school?',
'add_info': 'FOR ENUMERATOR: Please note any additional information that is
important for us',
'instanceID': 'instanceID',
'instanceName': 'instanceName',
'_v29': '_id',
'_v30': '_uuid',
'_v31': '_submission_time'}

```

[23]: df_4

```

[23]:
      start      today  team_lead  enum_name  \
0  2024-05-02T13:00:33.685Z  2.030227e+12      3      301
1  2024-04-22T10:27:49.144Z  2.029363e+12      5      501
2  2024-04-24T13:46:03.475Z  2.029536e+12      5      501
3  2024-04-24T11:27:07.968Z  2.029536e+12      5      501
4  2024-04-23T08:27:12.662Z  2.029450e+12      5      501
..      ...      ...      ...      ...
141  2024-04-26T08:17:30.031Z  2.029709e+12      4      401
142  2024-04-26T11:57:33.149Z  2.029709e+12      4      401
143  2024-04-29T08:28:06.303Z  2.029968e+12      4      401

```

144	2024-04-26T10:29:39.883Z	2.029709e+12	4	401
145	2024-04-25T13:22:11.645Z	2.029622e+12	4	401

	gps	_v1	_v2	_v3	_v4	\
0	17.3505286 -13.8823365 75.3 4.5	17.350529 -13.882336	75.3	4.500		
1	17.1219157 -14.0669924 76.0 4.88	17.121916 -14.066992	76.0	4.880		
2	17.0472362 -13.8998554 75.2 4.866	17.047236 -13.899855	75.2	4.866		
3	17.0289048 -13.9666087 74.1 4.7	17.028905 -13.966609	74.1	4.700		
4	17.097394 -13.8910092 86.2 4.671	17.097394 -13.891009	86.2	4.671		
..		
141	16.3740633 -13.1227434 54.6 5.0	16.374063 -13.122743	54.6	5.000		
142	16.4119209 -13.135757 90.3 4.883	16.411921 -13.135757	90.3	4.883		
143	16.4185709 -13.1366271 73.0 3.666	16.418571 -13.136627	73.0	3.666		
144	16.4066204 -13.1362544 80.5 4.78	16.406620 -13.136254	80.5	4.780		
145	16.158213 -12.7228823 62.5 4.966	16.158213 -12.722882	62.5	4.966		

	region	...	_v28	water_issues_other	water_on_property	\
0	1	...	NaN			1
1	1	...	0			1
2	1	...	0			1
3	1	...	NaN			1
4	1	...	1	Il n'ya pas de problÃme d'eau		1
..	
141	2	...	NaN			1
142	2	...	0			1
143	2	...	NaN			1
144	2	...	NaN			1
145	2	...	NaN			1

	water_walk	add_info	\
0	1 - pas de rÃefectoire \n- clÃture en mauvais Ã...		
1	1 Il y a un fut Ã eau oÃ les ÃolÃves boivent		
2	1 Il ya des lavabo qui ne sont plus utilisÃs c...		
3	1 Les latrines sont rarement utilisÃes par les ...		
4	1 L'Ãcole a le problÃme de rÃefectoire oÃ les...		
..	
141	1 Pas d'informations supplÃmentaires		
142	1 Pas d'informations supplÃmentaires		
143	1 Pas d'observation		
144	1 Manque de magasin pour la conservation des ali...		
145	1 Une fois sur place Ã sabahalla, on a constatÃ...		

	instanceID	instanceName	_v29	\
0	uuid:ab6c7b35-b8fd-4ba8-bb3f-30a3479bca76	1 - 2024-05-02	136299334	
1	uuid:d388ad9f-226a-4851-b4fe-ef97203fdee5	2 - 2024-04-22	135786482	
2	uuid:80cf7f21-0ca6-4bd5-bb32-b10bdbcfb033	3 - 2024-04-24	135940688	
3	uuid:b1dc2d4c-6798-4bac-8146-0c4c21692da0	4 - 2024-04-24	135940697	

```

4      uuid:1efdc50a-8d84-411d-99d3-453255f9156b      5 - 2024-04-23      135940701
..      ...
141    uuid:87bbb303-0386-4fb6-a5d1-890b0ae3f329      142 - 2024-04-26      135943890
142    uuid:fd493968-0c38-4f11-b441-86f53a150dc4      143 - 2024-04-26      135943941
143    uuid:2d906482-caa7-47d1-9821-3a393e56a833      144 - 2024-04-29      136096825
144    uuid:34ab1cad-bea6-4851-bd84-787d5742529c      145 - 2024-04-26      135943951
145    uuid:35ac61d5-17a0-4209-b23b-068d90daf01d      146 - 2024-04-25      135916800

```

```

                                _v30      _v31
0      ab6c7b35-b8fd-4ba8-bb3f-30a3479bca76      2.030281e+12
1      d388ad9f-226a-4851-b4fe-ef97203fdee5      2.029477e+12
2      80cf7f21-0ca6-4bd5-bb32-b10bdbcbf033      2.029763e+12
3      b1dc2d4c-6798-4bac-8146-0c4c21692da0      2.029763e+12
4      1efdc50a-8d84-411d-99d3-453255f9156b      2.029763e+12
..      ...
141    87bbb303-0386-4fb6-a5d1-890b0ae3f329      2.029769e+12
142    fd493968-0c38-4f11-b441-86f53a150dc4      2.029769e+12
143    2d906482-caa7-47d1-9821-3a393e56a833      2.030007e+12
144    34ab1cad-bea6-4851-bd84-787d5742529c      2.029769e+12
145    35ac61d5-17a0-4209-b23b-068d90daf01d      2.029694e+12

```

[146 rows x 343 columns]

Fusion des données

[]:

0.1 Données Élèves (df_student)

- **school_eat** : Lieu où l'élève mange à l'école (cantine, maison, autre).
- **school_food_like** : L'élève aime-t-il la nourriture servie à l'école ?
- **econ_hungry** : L'élève s'est-il endormi en ayant faim au cours des derniers jours ?
- **reading_activity** : Activités de lecture en classe pratiquées par l'élève (lecture en groupe, devoirs, jeux...).
- **reading_participate** : Participation à des activités extrascolaires liées à la lecture.
- **f5_wrong** : Nombre d'erreurs en lecture de mots (français).
- **f5_confirm**: 'Please confirm that the student read(français)
- **ar5_wrong** : Score total en lecture de mots (arabe).
- **ar_confirm**: 'Please confirm that the student read (arabe)
- **age** : Âge de l'élève.
- **region** : Région administrative de l'école.
- **commune** : Commune dans laquelle se situe l'école.

0.2 Données Enseignants (df_teacher)

- **absent** : Nombre de jours d'absence de l'enseignant durant le trimestre.
 - **absent_reason** : Raisons de l'absence de l'enseignant (maladie, conditions de travail, etc.).
 - **attent3** : Raisons perçues par l'enseignant du manque d'attention des élèves (fatigue, faim, distractions...).
 - **educ_mat** : L'enseignant a-t-il reçu du matériel pédagogique depuis 2020 ?
 - **educ_mat_what** : Type de matériel pédagogique reçu (livres, jeux, posters...).
 - **educ_mat_needs** : Évaluation de l'adéquation entre le matériel reçu et les besoins réels de l'enseignant.
 - **attent1** : Fréquence à laquelle l'enseignant interrompt la classe à cause d'un manque d'attention.
 - **attent2** : L'enseignant pense-t-il que l'attention des élèves a baissé cette année ?
-

0.3 Données Directeurs (df_director)

- **total_latrines** : Nombre total de latrines disponibles dans l'école.
 - **latrine_clean** : Niveau de propreté des latrines (évalué par le directeur).
 - **book_club** : Présence de clubs de lecture dans l'école.
 - **electronic_content** : Les enseignants utilisent-ils du contenu numérique dans leurs leçons ?
 - **program_received/New Arabic reading materials** : Distribution de nouveaux livres de lecture arabe dans le cadre du programme.
 - **attend_jan** : Taux moyen de présence des élèves en janvier.
 - **attend_feb** : Taux moyen de présence des élèves en février.
 - **attend_march** : Taux moyen de présence des élèves en mars.
 - **total_students** : Effectif total d'élèves dans l'école (à calculer si manquant).
 - **region** : Région de l'école.
 - **commune** : Commune de l'école.
-

[]:

la fusion des variables

[8]: `import pandas as pd`

```
# Étape 1 : Charger les deux fichiers nécessaires (élèves et enseignants)
df_teacher = pd.read_stata('/content/drive/MyDrive/DATA CHALLENGE/DATA_
↳CHALLENGE/Quant data before appending/teacher survey.dta')
```

```

df_student = pd.read_stata('/content/drive/MyDrive/DATA CHALLENGE/DATA_
↳CHALLENGE/Quant data before appending/egra and student survey.dta')

# Étape 2 : Définir les variables utiles par base
variables_student = [
    'school', 'region', 'commune',          # pour la fusion
    'econ_hungry',
    'f1_total',
    'f3_wrong',
    'ar3_total',
    'age'
]

variables_teacher = [
    'school', 'region', 'commune',          # pour la fusion
    'absent',
    'absent_reason',
    'attent3',
    'educ_mat',
    'educ_mat_what',
    'educ_mat_needs',
    'attent1',
    'attent2'
]

# Étape 3 : Extraire les colonnes nécessaires
df_student_clean = df_student[[col for col in variables_student if col in
↳df_student.columns]]
df_teacher_clean = df_teacher[[col for col in variables_teacher if col in
↳df_teacher.columns]]

# Étape 4 : Fusion des deux bases sur les clés communes
df_final = pd.merge(df_student_clean, df_teacher_clean, on=['school', 'region',
↳'commune'], how='right', suffixes=('', '_teacher'))

# Étape 5 : Sauvegarde du fichier fusionné
df_final.to_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right1.csv', index=False)

print(" Fichier fusionné enregistré avec succès : data_fusion_student_teacher.
↳csv")

```

<ipython-input-8-399b9caa251e>:4: UnicodeWarning:

One or more strings in the dta file could not be decoded using utf-8, and so the fallback encoding of latin-1 is being used. This can happen when a file has been incorrectly encoded by Stata or some other software. You should verify the string values returned are correct.


```
df_teacher = pd.read_stata('/content/drive/MyDrive/DATA CHALLENGE/DATA
CHALLENGE/Quant data before appending/teacher survey.dta')
```

Fichier fusionné enregistré avec succès : data_fusion_student_teacher.csv

```
[9]: data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right1.csv')
data
```

```
[9]:
```

	school	region	commune	econ_hungry	f1_total	\
0	El Wad Ghareg	Brakna	Aghchourguitt	Never	NaN	
1	El Wad Ghareg	Brakna	Aghchourguitt	Never	9.0	
2	El Wad Ghareg	Brakna	Aghchourguitt	NaN	NaN	
3	El Wad Ghareg	Brakna	Aghchourguitt	NaN	14.0	
4	El Wad Ghareg	Brakna	Aghchourguitt	NaN	NaN	
...	
5167	Sabahalla	Gorgol	Forum Legleita	Never	NaN	
5168	Sabahalla	Gorgol	Forum Legleita	Never	NaN	
5169	Sabahalla	Gorgol	Forum Legleita	Never	NaN	
5170	Sabahalla	Gorgol	Forum Legleita	Most days/always	NaN	
5171	Sabahalla	Gorgol	Forum Legleita	Never	NaN	

	f3_wrong	ar3_total	age	absent	absent_reason	attent3	educ_mat	\
0	NaN	NaN	6.0	0	NaN	1	Yes	
1	3.0	NaN	8.0	0	NaN	1	Yes	
2	NaN	50.0	NaN	0	NaN	1	Yes	
3	7.0	NaN	NaN	0	NaN	1	Yes	
4	NaN	15.0	NaN	0	NaN	1	Yes	
...	
5167	NaN	NaN	13.0	0	NaN	97 6	Yes	
5168	NaN	NaN	9.0	0	NaN	97 6	Yes	
5169	NaN	NaN	12.0	0	NaN	97 6	Yes	
5170	NaN	NaN	15.0	0	NaN	97 6	Yes	
5171	NaN	NaN	16.0	0	NaN	97 6	Yes	

	educ_mat_what	\
0	Des livres de lecture, en arabe et en français	
1	Des livres de lecture, en arabe et en français	
2	Des livres de lecture, en arabe et en français	
3	Des livres de lecture, en arabe et en français	
4	Des livres de lecture, en arabe et en français	
...	...	
5167	Guide de maître en arabe, livre troisième année...	
5168	Guide de maître en arabe, livre troisième année...	
5169	Guide de maître en arabe, livre troisième année...	
5170	Guide de maître en arabe, livre troisième année...	
5171	Guide de maître en arabe, livre troisième année...	

		educ_mat_needs		attent1	attent2
0	2	partially meets my needs		2 times	Yes
1	2	partially meets my needs		2 times	Yes
2	2	partially meets my needs		2 times	Yes
3	2	partially meets my needs		2 times	Yes
4	2	partially meets my needs		2 times	Yes
...	
5167	4	Completely meets my needs	More than 4 times	4 times	Yes
5168	4	Completely meets my needs	More than 4 times	4 times	Yes
5169	4	Completely meets my needs	More than 4 times	4 times	Yes
5170	4	Completely meets my needs	More than 4 times	4 times	Yes
5171	4	Completely meets my needs	More than 4 times	4 times	Yes

[5172 rows x 16 columns]

[]:

Nettoyage et traitement des valeurs manquantes

[]:

```
[14]: # Supprimer les lignes dupliquées (basé sur toutes les colonnes)
df_final = df_final.drop_duplicates()

# Optionnel : réinitialiser les index
df_final = df_final.reset_index(drop=True)

# Sauvegarde du fichier nettoyé
df_final.to_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv', index=False)

print(" Lignes dupliquées supprimées avec succès.")
df_final.to_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv', index=False)
data1 = df_teacher = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')
data1
```

Lignes dupliquées supprimées avec succès.

[14]:	school	region	commune	econ_hungry	f1_total	\
0	El Wad Ghareg	Brakna	Aghchourguitt	Never	NaN	
1	El Wad Ghareg	Brakna	Aghchourguitt	Never	9.0	
2	El Wad Ghareg	Brakna	Aghchourguitt	NaN	NaN	
3	El Wad Ghareg	Brakna	Aghchourguitt	NaN	14.0	
4	El Wad Ghareg	Brakna	Aghchourguitt	NaN	NaN	
...	
4757	Sabahalla	Gorgol	Forum Legleita	Never	NaN	

4758	Sabahalla	Gorgol	Forum Legleita	Never	NaN
4759	Sabahalla	Gorgol	Forum Legleita	Never	NaN
4760	Sabahalla	Gorgol	Forum Legleita	Most days/always	NaN
4761	Sabahalla	Gorgol	Forum Legleita	Never	NaN

	f3_wrong	ar3_total	age	absent	absent_reason	attent3	educ_mat	\
0	NaN	NaN	6.0	0	NaN	1	Yes	
1	3.0	NaN	8.0	0	NaN	1	Yes	
2	NaN	50.0	NaN	0	NaN	1	Yes	
3	7.0	NaN	NaN	0	NaN	1	Yes	
4	NaN	15.0	NaN	0	NaN	1	Yes	
...	
4757	NaN	NaN	13.0	0	NaN	97 6	Yes	
4758	NaN	NaN	9.0	0	NaN	97 6	Yes	
4759	NaN	NaN	12.0	0	NaN	97 6	Yes	
4760	NaN	NaN	15.0	0	NaN	97 6	Yes	
4761	NaN	NaN	16.0	0	NaN	97 6	Yes	

	educ_mat_what	\
0	Des livres de lecture, en arabe et en français	
1	Des livres de lecture, en arabe et en français	
2	Des livres de lecture, en arabe et en français	
3	Des livres de lecture, en arabe et en français	
4	Des livres de lecture, en arabe et en français	
...	...	
4757	Guide de maître en arabe, livre troisième année...	
4758	Guide de maître en arabe, livre troisième année...	
4759	Guide de maître en arabe, livre troisième année...	
4760	Guide de maître en arabe, livre troisième année...	
4761	Guide de maître en arabe, livre troisième année...	

	educ_mat_needs	attent1	attent2
0	2 partially meets my needs	2 times	Yes
1	2 partially meets my needs	2 times	Yes
2	2 partially meets my needs	2 times	Yes
3	2 partially meets my needs	2 times	Yes
4	2 partially meets my needs	2 times	Yes
...
4757	4 Completely meets my needs	More than 4 times	Yes
4758	4 Completely meets my needs	More than 4 times	Yes
4759	4 Completely meets my needs	More than 4 times	Yes
4760	4 Completely meets my needs	More than 4 times	Yes
4761	4 Completely meets my needs	More than 4 times	Yes

[4762 rows x 16 columns]

```
[15]: data_challenge = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')
data_challenge
```

```
[15]:
```

	school	region	commune	econ_hungry	f1_total	\
0	El Wad Ghareg	Brakna	Aghchourguitt	Never	NaN	
1	El Wad Ghareg	Brakna	Aghchourguitt	Never	9.0	
2	El Wad Ghareg	Brakna	Aghchourguitt	NaN	NaN	
3	El Wad Ghareg	Brakna	Aghchourguitt	NaN	14.0	
4	El Wad Ghareg	Brakna	Aghchourguitt	NaN	NaN	
...	
4757	Sabahalla	Gorgol	Forum Legleita	Never	NaN	
4758	Sabahalla	Gorgol	Forum Legleita	Never	NaN	
4759	Sabahalla	Gorgol	Forum Legleita	Never	NaN	
4760	Sabahalla	Gorgol	Forum Legleita	Most days/always	NaN	
4761	Sabahalla	Gorgol	Forum Legleita	Never	NaN	

	f3_wrong	ar3_total	age	absent	absent_reason	attent3	educ_mat	\
0	NaN	NaN	6.0	0	NaN	1	Yes	
1	3.0	NaN	8.0	0	NaN	1	Yes	
2	NaN	50.0	NaN	0	NaN	1	Yes	
3	7.0	NaN	NaN	0	NaN	1	Yes	
4	NaN	15.0	NaN	0	NaN	1	Yes	
...	
4757	NaN	NaN	13.0	0	NaN	97 6	Yes	
4758	NaN	NaN	9.0	0	NaN	97 6	Yes	
4759	NaN	NaN	12.0	0	NaN	97 6	Yes	
4760	NaN	NaN	15.0	0	NaN	97 6	Yes	
4761	NaN	NaN	16.0	0	NaN	97 6	Yes	

	educ_mat_what	\
0	Des livres de lecture, en arabe et en français	
1	Des livres de lecture, en arabe et en français	
2	Des livres de lecture, en arabe et en français	
3	Des livres de lecture, en arabe et en français	
4	Des livres de lecture, en arabe et en français	
...	...	
4757	Guide de maître en arabe, livre troisième année...	
4758	Guide de maître en arabe, livre troisième année...	
4759	Guide de maître en arabe, livre troisième année...	
4760	Guide de maître en arabe, livre troisième année...	
4761	Guide de maître en arabe, livre troisième année...	

	educ_mat_needs	attent1	attent2
0	2 partially meets my needs	2 times	Yes
1	2 partially meets my needs	2 times	Yes
2	2 partially meets my needs	2 times	Yes

3	2	partially meets my needs	2	times	Yes
4	2	partially meets my needs	2	times	Yes
...		
4757	4	Completely meets my needs	More than	4 times	Yes
4758	4	Completely meets my needs	More than	4 times	Yes
4759	4	Completely meets my needs	More than	4 times	Yes
4760	4	Completely meets my needs	More than	4 times	Yes
4761	4	Completely meets my needs	More than	4 times	Yes

[4762 rows x 16 columns]

```
[16]: # Afficher le type de chaque colonne dans le DataFrame
types_variables = data_challenge.dtypes

# Affichage propre avec tri alphabétique
print(" Types de chaque variable :\n")
print(types_variables.sort_index())
```

Types de chaque variable :

```
absent          int64
absent_reason   object
age             float64
ar3_total       float64
attent1         object
attent2         object
attent3         object
commune         object
econ_hungry     object
educ_mat        object
educ_mat_needs  object
educ_mat_what   object
f1_total        float64
f3_wrong        float64
region          object
school          object
dtype: object
```

```
[17]: # Afficher le nombre de valeurs manquantes pour chaque variable
missing_values = data_challenge.isnull().sum().sort_values(ascending=False)

# Affichage lisible
print(" Nombre de valeurs manquantes par variable :\n")
print(missing_values)
```

Nombre de valeurs manquantes par variable :

```
ar3_total      3656
```

```

f1_total      3407
f3_wrong      3407
absent_reason  2627
attent3       2413
econ_hungry   2036
age           2036
educ_mat_what 1567
region        0
school        0
commune       0
absent        0
educ_mat      0
educ_mat_needs 0
attent1       0
attent2       0
dtype: int64

```

les implutation

```

[18]: import pandas as pd
from sklearn.impute import SimpleImputer, KNNImputer
from sklearn.preprocessing import OrdinalEncoder
from sklearn.metrics import accuracy_score
import numpy as np

# Charger le fichier
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')

# Variables qualitatives à imputer
qual_vars = ['absent_reason', 'attent3', 'econ_hungry', 'educ_mat_what']

# Résultats finaux de comparaison
results = {}

# Pour chaque variable, simuler une perte de données, imputer, puis comparer
for col in qual_vars:
    col_data = data[[col]].dropna()
    if col_data.empty:
        continue

    # Encoder les catégories
    encoder = OrdinalEncoder()
    encoded = encoder.fit_transform(col_data)

    # Masquer artificiellement 20% des valeurs
    np.random.seed(42)

```

```

mask = np.random.rand(len(encoded)) < 0.2
original_values = encoded[mask].copy()
encoded[mask] = np.nan

# Imputation par mode
mode_imputer = SimpleImputer(strategy='most_frequent')
mode_imputed = mode_imputer.fit_transform(encoded)
acc_mode = accuracy_score(original_values, mode_imputed[mask])

# Imputation par constante (-1)
const_imputer = SimpleImputer(strategy='constant', fill_value=-1)
const_imputed = const_imputer.fit_transform(encoded)
acc_const = accuracy_score(original_values, const_imputed[mask])

# Imputation par KNN
knn_imputer = KNNImputer(n_neighbors=5)
knn_imputed = knn_imputer.fit_transform(encoded)
acc_knn = accuracy_score(original_values, np.round(knn_imputed[mask]))

# Stocker les résultats
results[col] = {
    'accuracy_mode': acc_mode,
    'accuracy_constant': acc_const,
    'accuracy_knn': acc_knn
}

# Résultats finaux
results_df = pd.DataFrame(results).T
print("Comparaison des méthodes d'imputation :")
print(results_df)

```

Comparaison des méthodes d'imputation :

	accuracy_mode	accuracy_constant	accuracy_knn
absent_reason	0.540659	0.0	0.035165
attent3	0.195171	0.0	0.004024
econ_hungry	0.514938	0.0	0.063269
educ_mat_what	0.147641	0.0	0.001522

```

[19]: from sklearn.impute import SimpleImputer

# Variables concernées
vars_to_impute = ['absent_reason', 'attent3', 'econ_hungry', 'educ_mat_what']

# Imputation par la modalité la plus fréquente
imputer_mode = SimpleImputer(strategy='most_frequent')

```

```

# Appliquer l'imputation
data[vars_to_impute] = imputer_mode.fit_transform(data[vars_to_impute])

# Sauvegarder les données si souhaité
data.to_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv', index=False)

print(" Imputation terminée. Fichier mis à jour :␣
↳data_fusion_student_teacher_final.csv")

```

Imputation terminée. Fichier mis à jour :
data_fusion_student_teacher_final.csv

```

[20]: # Imports nécessaires
import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.impute import SimpleImputer
from sklearn.metrics import mean_squared_error

# Charger le fichier
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')

# Extraire "age" + 3 variables corrélées pour la régression
age_data = data[['age', 'f1_total', 'f3_wrong', 'ar3_total']].copy()
age_data = age_data.dropna(how='all')

# Masquer 20% des âges non manquants (pour tester l'imputation)
np.random.seed(42)
mask = (age_data['age'].notna()) & (np.random.rand(len(age_data)) < 0.2)
true_age = age_data.loc[mask, 'age'].copy()
age_data.loc[mask, 'age'] = np.nan

# Imputation 1: Médiane
imp_median = SimpleImputer(strategy='median')
age_median = imp_median.fit_transform(age_data)
mse_median = mean_squared_error(true_age, age_median[mask, 0])

# Imputation 2: Moyenne
imp_mean = SimpleImputer(strategy='mean')
age_mean = imp_mean.fit_transform(age_data)
mse_mean = mean_squared_error(true_age, age_mean[mask, 0])

# Imputation 3: Mode
imp_mode = SimpleImputer(strategy='most_frequent')
age_mode = imp_mode.fit_transform(age_data)

```



```

mse_mode = mean_squared_error(true_age, age_mode[mask, 0])

# Imputation 4: Régression linéaire
train = age_data[age_data['age'].notna()]
test = age_data[age_data['age'].isna()]

X_train = train[['f1_total', 'f3_wrong', 'ar3_total']].fillna(0)
y_train = train['age']
X_test = test[['f1_total', 'f3_wrong', 'ar3_total']].fillna(0)

model = LinearRegression()
model.fit(X_train, y_train)
age_pred = model.predict(X_test)

# Évaluer uniquement les prédictions sur les index artificiellement masqués
index_masked = age_data[mask].index
age_pred_masked = pd.Series(age_pred, index=test.index).loc[index_masked]
mse_reg = mean_squared_error(true_age, age_pred_masked)

# Résumé des résultats
mse_results = pd.DataFrame({
    'Méthode': ['Médiane', 'Moyenne', 'Mode', 'Régression Linéaire'],
    'MSE': [mse_median, mse_mean, mse_mode, mse_reg]
})

print("Comparaison des méthodes d'imputation :")
print(mse_results)

```

Comparaison des méthodes d'imputation :

	Méthode	MSE
0	Médiane	25.560477
1	Moyenne	25.299614
2	Mode	25.560477
3	Régression Linéaire	24.779245

```

[57]: import pandas as pd
import numpy as np
from sklearn.linear_model import LinearRegression

# Charger les données
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')

# Préparer les colonnes nécessaires pour la régression
features = ['f1_total', 'f3_wrong', 'ar3_total']
df_age = data[['age'] + features].copy()

```

```

# Définir les lignes d'entraînement et de prédiction
train = df_age[df_age['age'].notna()]
test = df_age[df_age['age'].isna()]

X_train = train[features].fillna(0)
y_train = train['age']
X_test = test[features].fillna(0)

# Entraîner le modèle
model = LinearRegression()
model.fit(X_train, y_train)

# Prédire les âges manquants
age_pred = model.predict(X_test)

# Appliquer l'imputation dans la base d'origine
data.loc[data['age'].isna(), 'age'] = age_pred

# (Optionnel) Sauvegarder les données mises à jour
data.to_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv', index=False)

print(" Imputation terminée avec la régression linéaire.")

```

Imputation terminée avec la régression linéaire.

```

[58]: df_check = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')
print(df_check['age'].isna().sum())

```

0

```

[59]: # Afficher le nombre de valeurs manquantes pour chaque variable
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')
missing_values = data.isnull().sum().sort_values(ascending=False)

# Affichage lisible
print(" Nombre de valeurs manquantes par variable :\n")
print(missing_values)

```

Nombre de valeurs manquantes par variable :

ar3_total	3656
f1_total	3527
f3_wrong	3407
school	0
econ_hungry	0
commune	0

```

region          0
age             0
absent          0
absent_reason   0
attent3         0
educ_mat        0
educ_mat_what   0
educ_mat_needs  0
attent1         0
attent2         0
dtype: int64

```

Cartoghaghie

```

[29]: import pandas as pd
import folium
from folium.plugins import MarkerCluster

# Charger les données
df = pd.read_stata('/content/drive/MyDrive/DATA CHALLENGE/DATA CHALLENGE/Quant_
↳data before appending/egra and student survey.dta')

# Filtrer les colonnes nécessaires
cols = ['school', 'region', 'commune', '_v1', '_v2']
df = df[cols].dropna(subset=['_v1', '_v2']).drop_duplicates()

# Créer la carte
mauritania_map = folium.Map(location=[df['_v1'].mean(), df['_v2'].mean()],
↳zoom_start=6)

# Ajouter les marqueurs
cluster = MarkerCluster().add_to(mauritania_map)
for _, row in df.iterrows():
    folium.Marker(
        location=[row['_v1'], row['_v2']],
        popup=f"École: {row['school']}<br>Région: {row['region']}<br>Commune:
↳{row['commune']}"
    ).add_to(cluster)

# Afficher dans Colab
mauritania_map

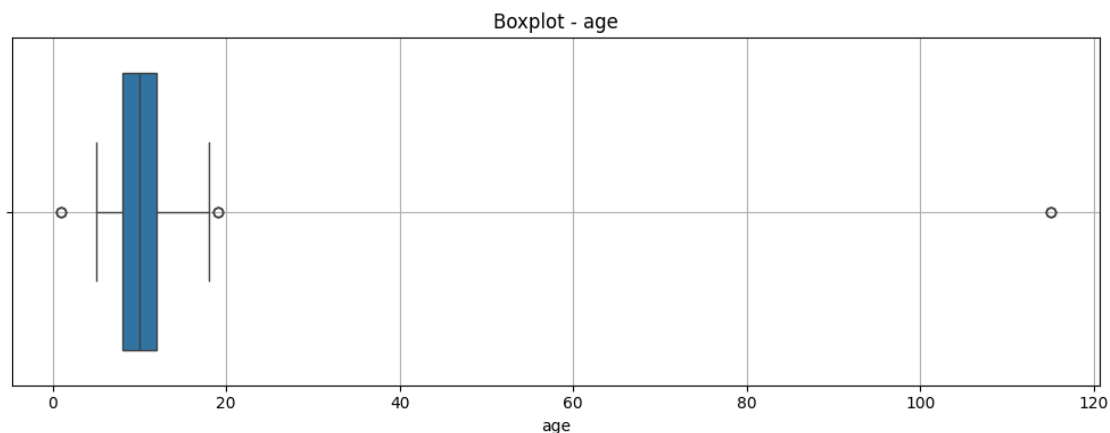
```

[29]: <folium.folium.Map at 0x7999fc06a490>

ANALYSE EXPLORATOIRE

0.4 Boxplot - Age

```
[55]: import matplotlib.pyplot as plt
import seaborn as sns
plt.figure(figsize=(10, 4))
sns.boxplot(x=data_challenge['age'])
plt.title('Boxplot - age')
plt.grid(True)
plt.tight_layout()
plt.show()
```



Interpretation

La variable “âge” montre que la majorité des élèves ont un âge compris entre **8 et 15 ans**, ce qui correspond logiquement à la tranche d’âge attendue pour des élèves du primaire. Toutefois, on observe des **valeurs aberrantes très élevées**, dépassant parfois les **100 ans**, ce qui est manifestement irréaliste dans ce contexte. Ces valeurs extrêmes sont probablement dues à des **erreurs de saisie** lors de la collecte des données ou à des **problèmes survenus lors de l’imputation** des valeurs manquantes. Il est donc fortement recommandé de **corriger ou d’exclure ces âges anormaux** avant d’entreprendre des analyses statistiques ou des modélisations, afin d’assurer la **fiabilité des résultats**.

Correction de Valeurs aberrantes

```
[33]: import pandas as pd

# Charger le fichier
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')

# Détection et remplacement des outliers de l'âge
q1 = data['age'].quantile(0.25)
q3 = data['age'].quantile(0.75)
```

```

iqr = q3 - q1

# Bornes raisonnables (ajustables)
lower_bound = max(6, q1 - 1.5 * iqr)
upper_bound = min(20, q3 + 1.5 * iqr)

# Affichage avant
print(f"Âges avant traitement : min = {data['age'].min()}, max = {data['age'].max()}")

# Remplacement des outliers par NaN
data.loc[(data['age'] < lower_bound) | (data['age'] > upper_bound), 'age'] = pd.NA

# Option : vous pouvez ensuite réimputer les valeurs manquantes avec la méthode choisie (ex. régression)
# ...

# Sauvegarde dans le même fichier ou un nouveau
data.to_csv('/content/drive/MyDrive/DATA CHALLENGE/data_fusion_student_teacher_final_right12.csv', index=False)

print(" Valeurs aberrantes de l'âge corrigées avec succès.")

```

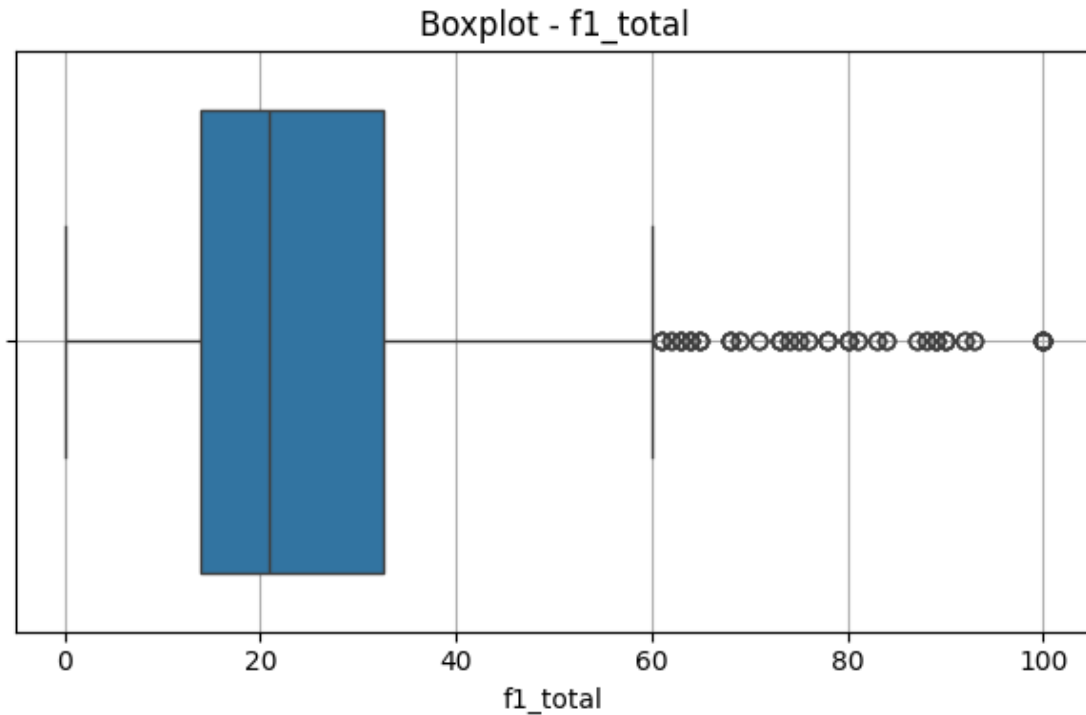
Âges avant traitement : min = 1.0, max = 115.0
 Valeurs aberrantes de l'âge corrigées avec succès.

0.5 Boxplot - f1_total (Score de lecture de lettres en français)

```

[37]: plt.figure(figsize=(6, 4))
sns.boxplot(x=data_challenge['f1_total'])
plt.title('Boxplot - f1_total')
plt.grid(True)
plt.tight_layout()
plt.show()

```



Intrepretation

Le boxplot de la variable `f1_total`, qui indique le score de lecture de lettres en français, révèle que les scores sont majoritairement concentrés entre 10 et 35 lettres lues, avec une médiane située autour de 25.

Bien que les valeurs les plus élevées du score `f1_total` apparaissent comme des outliers dans le boxplot, elles ne doivent pas être considérées comme des valeurs aberrantes au sens pédagogique. En effet, ces scores reflètent probablement des élèves ayant un bon niveau de lecture les lettres en français, ce qui peut s'expliquer par des différences d'environnement, de ressources ou de pratiques pédagogiques.

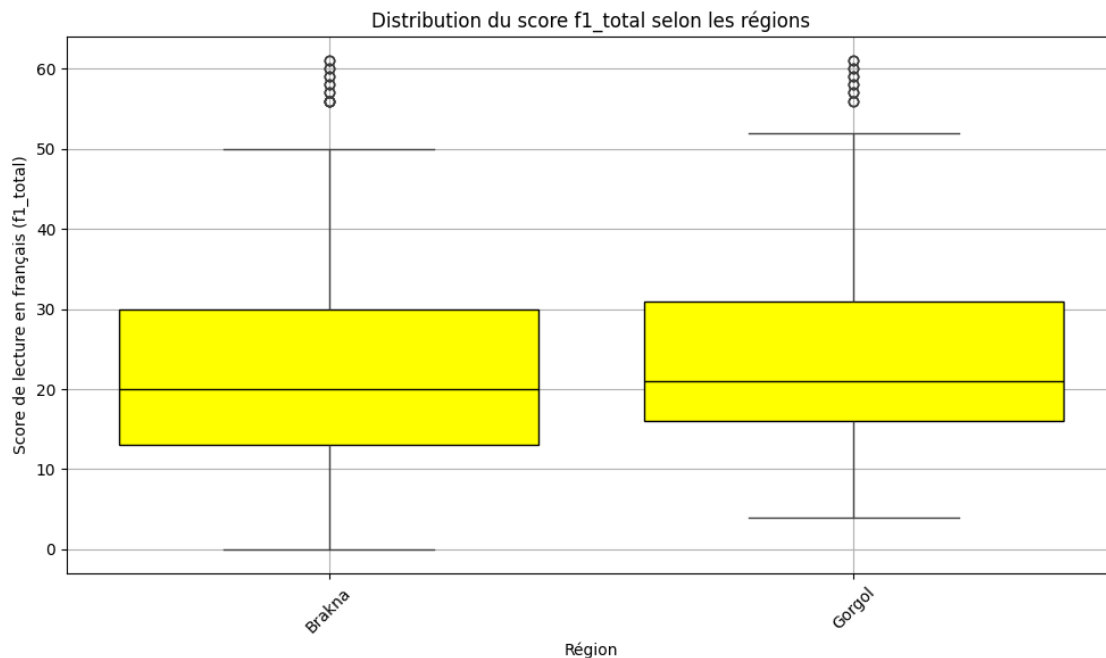
Il est donc pertinent de **les conserver** pour ne pas fausser l'analyse de la performance réelle des élèves et permettre une meilleure compréhension des écarts entre les apprenants.

```
[54]: import seaborn as sns
import matplotlib.pyplot as plt
import pandas as pd

# Charger les données
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final.csv')

# Nettoyage des données : garder seulement les valeurs non manquantes
df_plot = data[['f1_total', 'region']].dropna()
```

```
# Créer le boxplot avec couleur jaune
plt.figure(figsize=(10, 6))
sns.boxplot(
    x='region',
    y='f1_total',
    data=df_plot,
    boxprops=dict(facecolor='yellow', edgecolor='black'),
    medianprops=dict(color='black')
)
plt.title('Distribution du score f1_total selon les régions')
plt.xlabel('Région')
plt.ylabel('Score de lecture en français (f1_total)')
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
plt.show()
```



Intrepretation

L'analyse des distributions du score `f1_total` montre que la région du **Gorgol** présente une **médiane légèrement plus élevée** que celle de **Brakna**, suggérant que les élèves y lisent généralement **plus de lettres en français**.

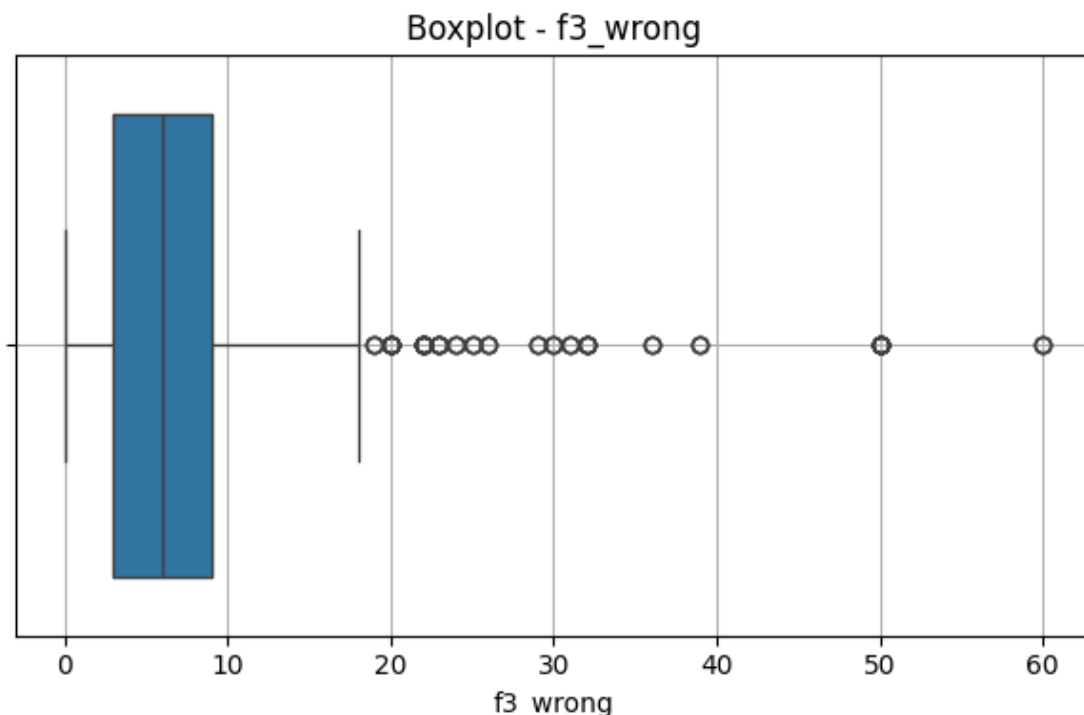
L'analyse du score de lecture en français (`f1_total`) entre les régions de Brakna et de Gorgol révèle que les élèves du Gorgol ont en moyenne des performances légèrement meilleures. Cette différence dans les scores suggère que les enseignants de français dans la région du Gorgol pourraient être

plus efficaces ou bénéficier de meilleures conditions pédagogiques. Elle peut aussi refléter une exposition plus forte à la langue française dans cette région, ce qui faciliterait l'apprentissage des élèves. Par conséquent, cette tendance soutient l'idée que les élèves du Gorgol ont un meilleur encadrement ou un environnement linguistique plus favorable à l'apprentissage du français que ceux de Brakna, ce qui est un élément important à considérer dans l'étude de l'impact des enseignants sur la performance scolaire.

Ces résultats soutiennent l'**hypothèse selon laquelle les enseignants de français dans la région du Gorgol seraient plus efficaces ou mieux formés** que ceux de Brakna. Cela pourrait également indiquer que les élèves du Gorgol **maîtrisent mieux le français**, ce qui favoriserait leur apprentissage et leur attention en classe. Il serait pertinent de croiser ces résultats avec d'autres indicateurs (matériel pédagogique, assiduité des enseignants, participation aux activités de lecture, etc.) pour confirmer cette tendance.

0.6 Boxplot - f3_wrong (Erreurs de lecture de mots en français)

```
[50]: plt.figure(figsize=(6, 4))
sns.boxplot(x=data_challenge['f3_wrong'])
plt.title('Boxplot - f3_wrong')
plt.grid(True)
plt.tight_layout()
plt.show()
```



0.7 interpretation

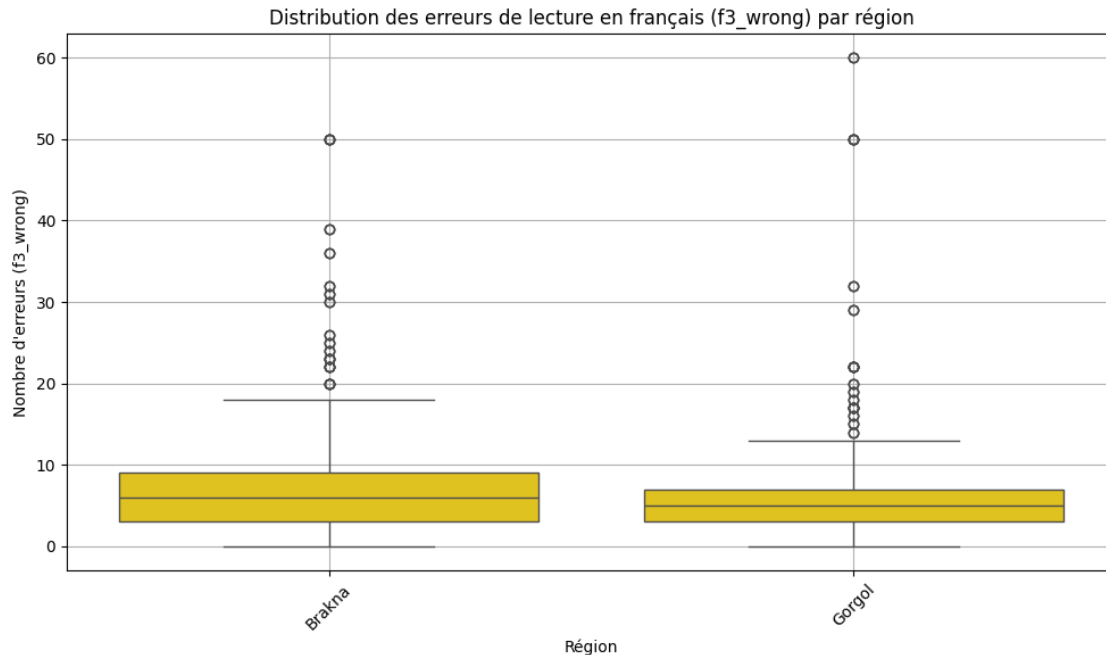
Le boxplot de la variable `f3_wrong`, représentant le nombre d'erreurs de lecture de mots en français, montre que la majorité des élèves font entre 0 et 10 erreurs, avec une médiane située autour de 5 erreurs. Cependant, on observe quelques valeurs aberrantes, notamment des élèves ayant commis plus de 20 erreurs, ce qui peut refléter une grande difficulté en lecture. Cette variable est donc particulièrement pertinente pour identifier les élèves en difficulté de compréhension ou d'apprentissage.

```
[53]: import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Chargement des données
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')

# Filtrer uniquement les lignes avec des valeurs valides
df = data[['region', 'f3_wrong']].dropna()

# Création du boxplot
plt.figure(figsize=(10, 6))
sns.boxplot(data=df, x='region', y='f3_wrong', color='gold')
plt.title("Distribution des erreurs de lecture en français (f3_wrong) par_
↳région")
plt.xlabel("Région")
plt.ylabel("Nombre d'erreurs (f3_wrong)")
plt.xticks(rotation=45)
plt.grid(True)
plt.tight_layout()
plt.show()
```



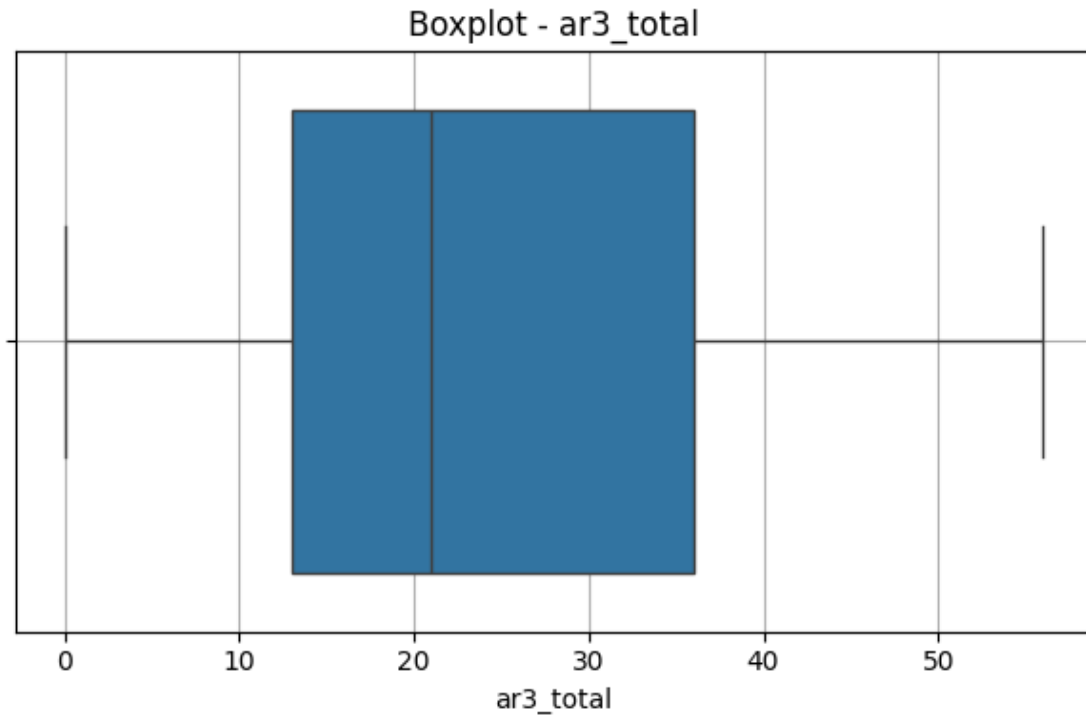
Intrepretation

L'analyse graphique des erreurs de lecture de mots en français (f3_wrong) montre une différence notable entre les élèves des régions de Brakna et de Gorgol. Dans la région de Brakna, les élèves commettent davantage d'erreurs, comme le suggère une médiane plus élevée et une plus grande dispersion des résultats, avec de nombreuses valeurs extrêmes. À l'inverse, les élèves de Gorgol présentent un nombre d'erreurs plus faible, avec une distribution plus resserrée et une médiane inférieure.

Cette disparité peut être interprétée comme un reflet des différences dans la qualité de l'enseignement du français entre les deux régions. Elle semble appuyer l'hypothèse selon laquelle les enseignants de Gorgol sont mieux formés, ou disposent de meilleures ressources pédagogiques, ce qui se traduit par une meilleure performance des élèves. Par ailleurs, il est possible que les élèves de Gorgol soient plus exposés à la langue française dans leur environnement quotidien, facilitant ainsi leur apprentissage.

En résumé, cette analyse renforce la problématique initiale en suggérant que les conditions d'enseignement du français, ainsi que les compétences des enseignants, varient selon les régions et influencent directement les performances des élèves.

```
[83]: plt.figure(figsize=(6, 4))
sns.boxplot(x=data['ar3_total'])
plt.title('Boxplot - ar3_total')
plt.grid(True)
plt.tight_layout()
plt.show()
```

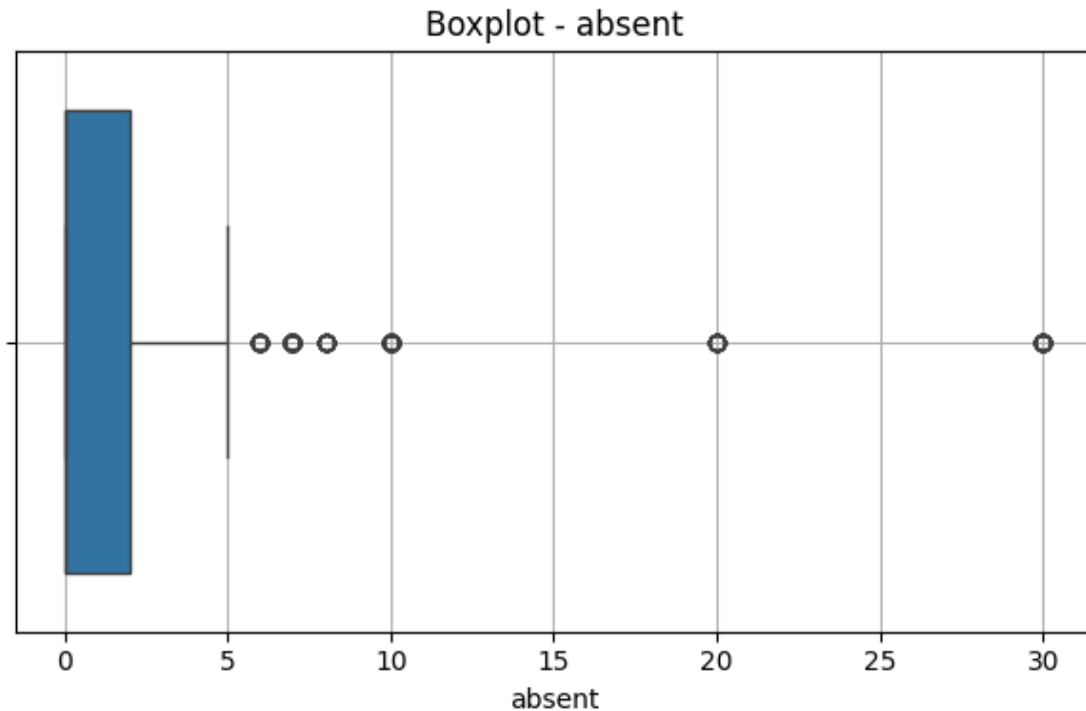


0.8 Boxplot - ar3_total (Score de lecture de mots en arabe)

Le boxplot de la variable `ar3_total`, qui mesure le score de lecture de mots en arabe, montre une distribution centrée entre 10 et 40, avec une médiane se situant approximativement entre 25 et 30. Cette répartition indique que la majorité des élèves obtiennent des scores dans cette plage, ce qui reflète une performance moyenne à satisfaisante. On observe très peu de valeurs aberrantes, ce qui témoigne d'une distribution assez régulière et d'une bonne qualité des données. Ainsi, cette variable peut être considérée comme fiable et bien mesurée pour évaluer les compétences de lecture en arabe des élèves.

0.9 Variable : absent (nombre de jours d'absence de l'enseignant)

```
[86]: plt.figure(figsize=(6, 4))
sns.boxplot(x=data['absent'])
plt.title('Boxplot - absent')
plt.grid(True)
plt.tight_layout()
plt.show()
```



0.10 interpretation

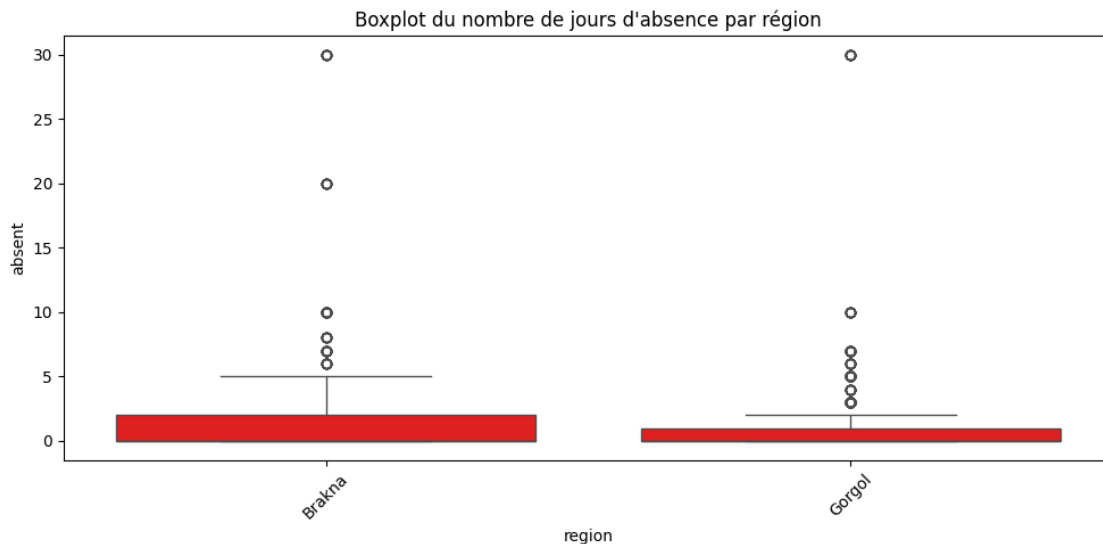
Le boxplot de la variable **absent**, qui représente le nombre de jours d'absence des enseignants, révèle une distribution très asymétrique. La majorité des enseignants sont absents entre 0 et 5 jours, ce qui montre que l'absentéisme est globalement faible. La médiane, située autour de 1 jour, confirme que plus de la moitié des enseignants ont une très faible fréquence d'absence. Toutefois, quelques observations isolées dépassent largement cette tendance, avec des absences allant jusqu'à 10, 20 voire 30 jours. Ces valeurs sont considérées comme des valeurs aberrantes. Elles pourraient correspondre à des cas particuliers (maladie prolongée, congé exceptionnel, ou saisie erronée) et mériteraient d'être examinées ou corrigées si elles biaisent l'analyse. Ce type de distribution suggère qu'une transformation ou une approche robuste pourrait être envisagée pour intégrer cette variable dans une analyse statistique plus avancée.

```
[84]: import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Charger les données depuis le fichier fusionné utilisé dans l'analyse
data = pd.read_csv("/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv")

# Création des boxplots par région, commune et école
plt.figure(figsize=(10, 5))
```

```
sns.boxplot(x="region", y="absent", data=data, color='red')
plt.title("Boxplot du nombre de jours d'absence par région")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



0.11 Interprétation du boxplot des absences par région

Le graphique montre la distribution du nombre de jours d'absence des enseignants dans les régions de Brakna et Gorgol. On observe que la **médiane** des jours d'absence est **plus élevée à Brakna** qu'à Gorgol, avec également une plus grande dispersion et plusieurs **valeurs extrêmes** (jusqu'à 30 jours).

Cela suggère que les enseignants de Brakna sont **plus fréquemment absents** que ceux de Gorgol, ce qui pourrait être le signe de **conditions de travail plus difficiles**, de **problèmes de santé plus fréquents**, ou d'un **manque d'encadrement et de suivi administratif**.

0.11.1 Conclusion

Ces résultats renforcent l'idée que **la qualité de vie et les conditions professionnelles des enseignants sont probablement meilleures à Gorgol** qu'à Brakna. Cette différence peut impacter directement la qualité de l'enseignement dispensé, notamment en matière de régularité et de suivi pédagogique auprès des élèves.

```
[88]: import pandas as pd

# Charger les données
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final.csv')
```

```

# Calculer la moyenne d'absences par école
moyennes_absence = data.groupby('school')['absent'].mean().
    ↪sort_values(ascending=False).head(20)

# Extraire les écoles concernées
top_ecoles = moyennes_absence.index.tolist()

# Associer école - région
regions_top = data[data['school'].isin(top_ecoles)][['school', 'region']].
    ↪drop_duplicates()

# Fusionner avec les moyennes
df_final = pd.merge(
    moyennes_absence.reset_index(name='absent_moyen'),
    regions_top,
    on='school',
    how='left'
)

print(df_final)

```

	school	absent_moyen	region
0	Garlol	11.333333	Brakna
1	Zahra	10.333333	Gorgol
2	Weindy Md Ethmane	6.666667	Brakna
3	Darsalam	4.333333	Brakna
4	T. Cive 1	4.000000	Gorgol
5	Mabrouk Pk18	4.000000	Brakna
6	Eronne	4.000000	Gorgol
7	Thenel	4.000000	Brakna
8	Elwihda	3.666667	Brakna
9	Agueda	3.666667	Brakna
10	Bassinguedi	3.333333	Brakna
11	Dar El Avia	3.333333	Brakna
12	Nejah	3.333333	Gorgol
13	Hamdallaye	3.000000	Brakna
14	Darsalam3	3.000000	Gorgol
15	Aleg 6	2.666667	Brakna
16	Advjeijir	2.666667	Brakna
17	Tadioukel	2.666667	Brakna
18	Lejouwad	2.666667	Brakna
19	Debaye Hijaj	2.333333	Brakna

```

[89]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

```

```

# Charger les données
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')

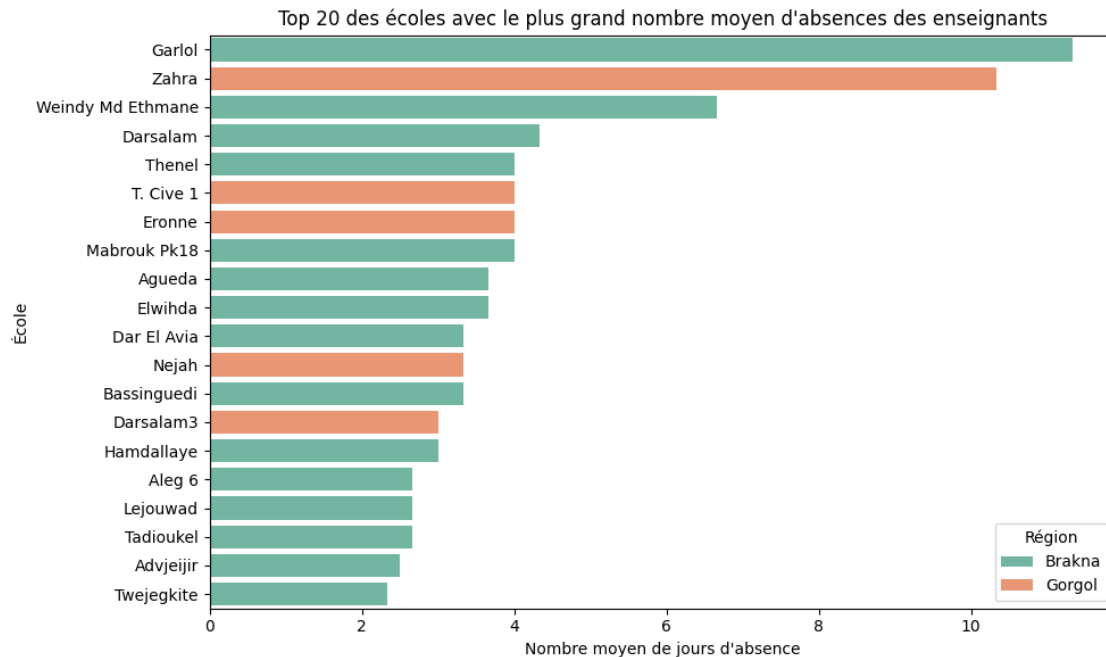
# Calculer la moyenne d'absences par école
moyennes_absence = data.groupby('school')['absent'].mean().
↳sort_values(ascending=False).head(20)
top_ecoles = moyennes_absence.index.tolist()

# Associer école à la région
regions_top = data[data['school'].isin(top_ecoles)][['school', 'region']].
↳drop_duplicates()

# Fusionner avec les moyennes
df_final = pd.merge(
    moyennes_absence.reset_index(name='absent_moyen'),
    regions_top,
    on='school',
    how='left'
)

# Tracer le graphique avec couleurs selon la région
plt.figure(figsize=(10, 6))
sns.barplot(data=df_final, y='school', x='absent_moyen', hue='region',
↳palette='Set2')
plt.xlabel("Nombre moyen de jours d'absence")
plt.ylabel("École")
plt.title("Top 20 des écoles avec le plus grand nombre moyen d'absences des_
↳enseignants")
plt.legend(title="Région")
plt.tight_layout()
plt.show()

```



0.11.2 Interprétation du top 20 des écoles avec le plus grand nombre moyen d'absences

Le graphique met en évidence que la majorité des écoles ayant le plus fort taux d'absences moyennes des enseignants se trouvent dans la région du **Brakna**. Cette surreprésentation peut s'expliquer par des **contraintes géographiques et sociales** : plusieurs de ces établissements sont situés dans des **communes reculées**, éloignées des centres urbains et dépourvues d'infrastructures de base. Dans ces conditions, les enseignants affectés dans ces zones rencontrent souvent des **difficultés à résider durablement**, ce qui impacte directement leur présence régulière en classe.

À l'inverse, dans la région du **Gorgol**, les communes sont généralement **plus proches des centres urbains**, ce qui facilite l'accessibilité et améliore les conditions de vie des enseignants. Cela peut expliquer pourquoi cette région est moins représentée dans ce classement et présente des **absences globalement plus faibles**.

Cette analyse souligne donc que **la qualité de vie locale et l'accessibilité des écoles** jouent un rôle important dans l'assiduité des enseignants, et par conséquent, dans la qualité de l'enseignement délivré.

[]:

Comparaison régionale

```
[90]: import pandas as pd
import matplotlib.pyplot as plt

# Charger les données
```



```

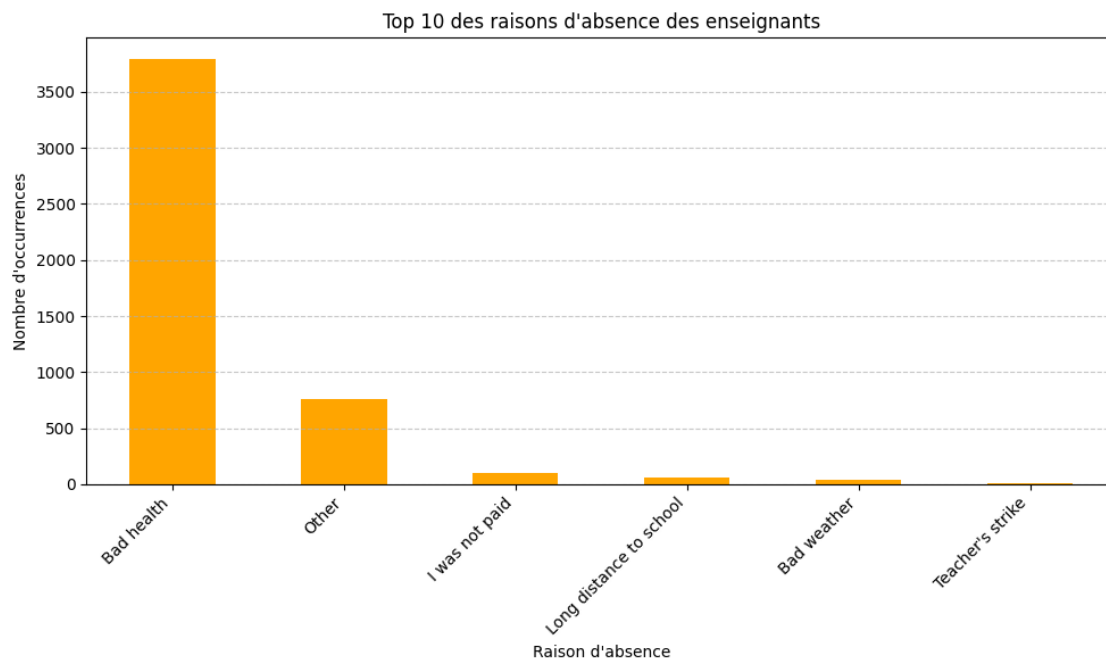
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')

# Compter les raisons d'absence
absent_reason_counts = data['absent_reason'].value_counts().head(10)

# Tracer le graphique avec axes inversés
plt.figure(figsize=(10, 6))
absent_reason_counts.plot(kind='bar', color='orange')

plt.title("Top 10 des raisons d'absence des enseignants")
plt.ylabel("Nombre d'occurrences")
plt.xlabel("Raison d'absence")
plt.xticks(rotation=45, ha='right')
plt.grid(True, axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()

```



0.11.3 Interprétation des raisons d'absence des enseignants

L'analyse des données révèle que la cause la plus fréquente d'absence des enseignants est liée à des problèmes de santé, représentant de loin la majorité des cas. Cette prédominance des absences pour raisons médicales souligne un problème structurel dans la qualité de vie et les conditions de travail du personnel éducatif. Un environnement professionnel difficile, combiné à un accès limité aux soins, peut conduire à une fréquence élevée de maladies ou à l'aggravation de troubles mineurs, ce qui augmente l'absentéisme.

Cette situation a un impact direct sur la qualité de l'enseignement, car l'absence répétée des enseignants interrompt la continuité pédagogique et compromet la progression des élèves. Par conséquent, ces résultats appuient la problématique selon laquelle les mauvaises conditions de travail, notamment liées à la santé et à l'hygiène, jouent un rôle central dans la performance scolaire. Investir dans l'amélioration de la santé et du bien-être des enseignants pourrait donc constituer un levier essentiel pour renforcer la qualité de l'enseignement et la réussite des élèves.

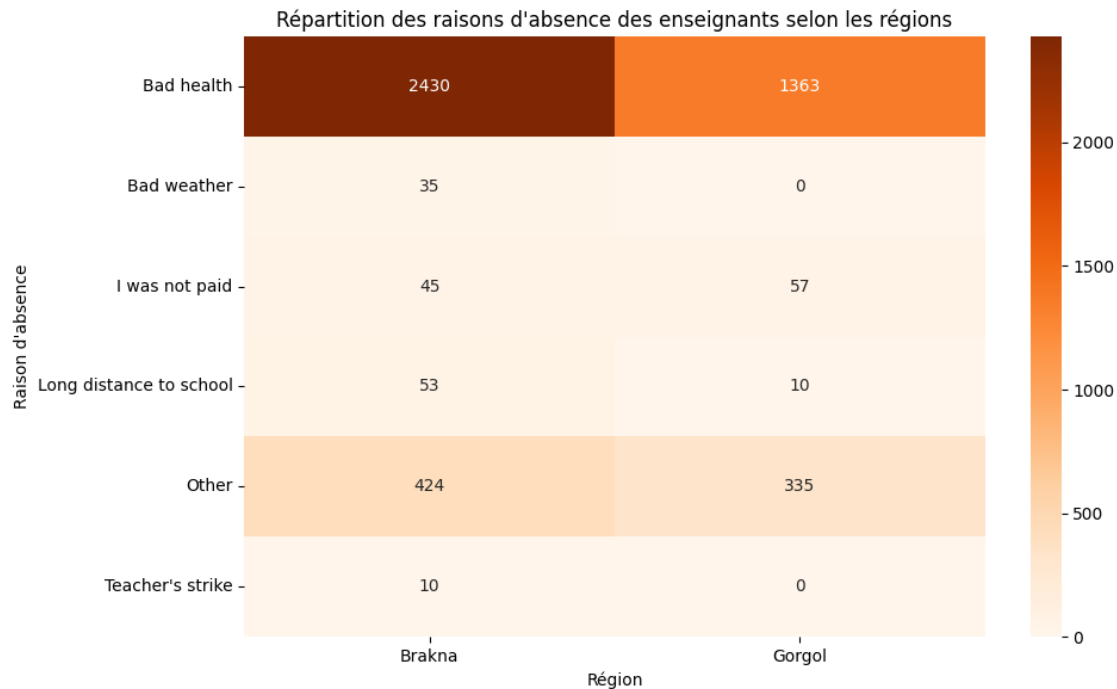
```
[81]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Recharger les données fusionnées
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')

# Nettoyage de la variable 'absent_reason'
df = data[['region', 'absent_reason']].dropna()

# Création du tableau croisé
pivot_table = df.pivot_table(index='absent_reason', columns='region',
↳aggfunc=len, fill_value=0)

# Tracer la heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(pivot_table, annot=True, fmt='d', cmap='Oranges')
plt.title("Répartition des raisons d'absence des enseignants selon les régions")
plt.xlabel("Région")
plt.ylabel("Raison d'absence")
plt.tight_layout()
plt.show()
```



0.11.4 Interprétation et conclusion

L'analyse des raisons d'absence des enseignants selon les régions révèle une disparité significative entre le **Brakna** et le **Gorgol**. On observe que **la mauvaise santé ("Bad health")** est de loin **la principale cause d'absence** dans les deux régions, mais elle est **beaucoup plus fréquente à Brakna** (2430 cas contre 1363 à Gorgol). Cela peut indiquer que les enseignants du Brakna sont confrontés à **des conditions de vie plus difficiles ou un accès limité aux soins de santé**, ce qui affecte directement leur présence à l'école.

Par ailleurs, des causes comme **"Bad weather"**, **"Long distance to school"** ou **"Teacher's strike"** apparaissent exclusivement ou plus fréquemment dans le Brakna, tandis que le **Gorgol affiche une meilleure stabilité sur ces aspects**.

On peut donc en conclure que **la qualité de vie des enseignants semble meilleure à Gorgol**. Cela pourrait expliquer en partie **les meilleurs scores des élèves en lecture dans cette région**, comme cela a été observé dans vos analyses précédentes. Un environnement de travail plus favorable, combiné à une présence plus régulière des enseignants, semble donc **favoriser la performance éducative** dans le Gorgol, renforçant l'idée d'un lien direc

```
[73]: import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

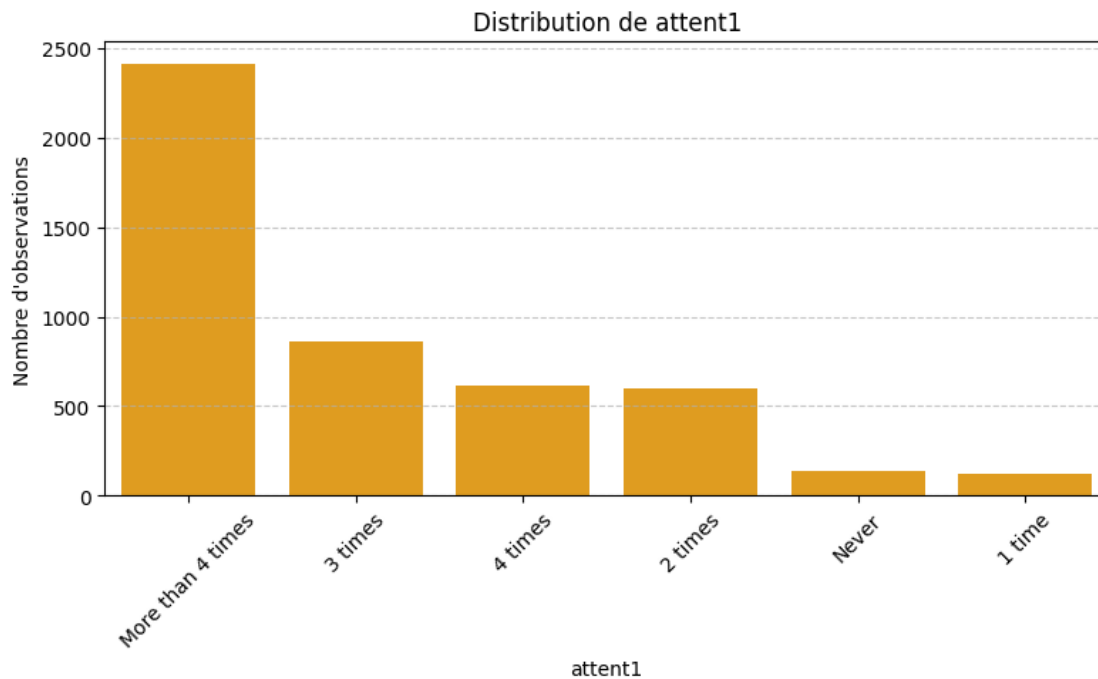
# Charger les données
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')
```

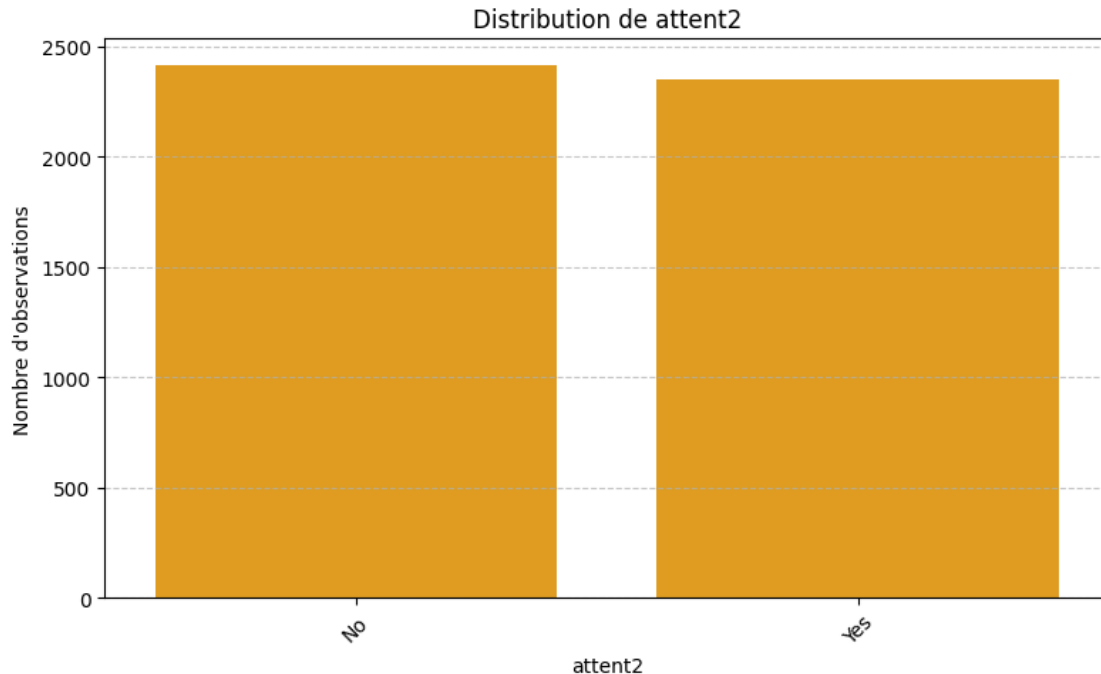
```

# Variables à analyser
variables_attention = ['attent1', 'attent2']

# Création d'un graphe pour chaque variable
for var in variables_attention:
    plt.figure(figsize=(8, 5))
    sns.countplot(data=data, x=var, order=data[var].value_counts().index,
        color='orange')
    plt.title(f'Distribution de {var}')
    plt.xlabel(var)
    plt.ylabel("Nombre d'observations")
    plt.xticks(rotation=45)
    plt.grid(axis='y', linestyle='--', alpha=0.7)
    plt.tight_layout()
    plt.show()

```





L'analyse des variables `attent1` et `attent2`, relatives à l'attention des élèves en classe, révèle des éléments préoccupants qui confirment notre problématique initiale sur l'impact de la qualité de l'enseignement sur la concentration des élèves. En effet, plus de 2600 enseignants déclarent interrompre le cours "plus de 4 fois" à cause du manque d'attention (`attent1`), ce qui indique une fréquence d'interruptions anormalement élevée. Cela suggère que les élèves éprouvent régulièrement des difficultés à suivre les leçons, ce qui nuit fortement à la continuité pédagogique. De plus, la variable `attent2` montre une répartition presque égale entre les enseignants affirmant que l'attention a baissé cette année (réponse "Yes") et ceux qui ne le pensent pas ("No"), ce qui souligne une tendance perçue vers la dégradation de la concentration en classe.

Ces constats confirment que les conditions actuelles (qu'elles soient liées à l'état de santé, la fatigue, les absences des enseignants ou les facteurs socio-économiques) affectent directement l'attention des élèves. Ils renforcent donc l'idée que des actions ciblées sur l'environnement d'apprentissage, la formation et le bien-être des enseignants sont nécessaires pour améliorer l'attention en classe et, in fine, la performance scolaire.

```
[72]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from collections import Counter

# Charger les données
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')
```

```

# Nettoyer et compter chaque code séparément (ex : "1 2 3" devient ['1', '2', '3'])
attent3_cleaned = data['attent3'].dropna().astype(str).str.split()
all_codes = [code for sublist in attent3_cleaned for code in sublist]

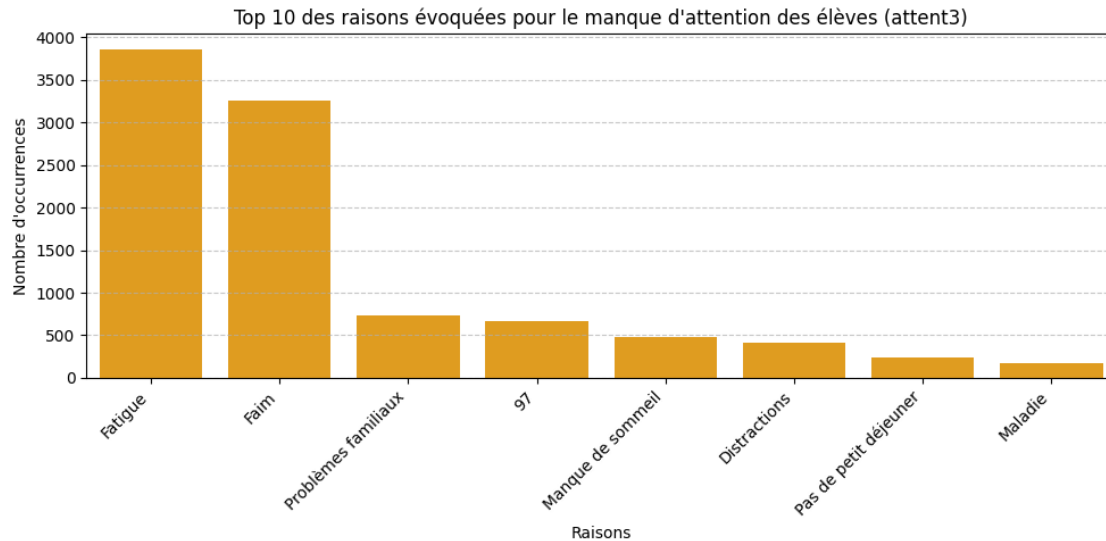
# Compter les fréquences
code_counts = pd.DataFrame(Counter(all_codes).items(), columns=['Code', 'Occurrences']).sort_values(by='Occurrences', ascending=False)

# (Optionnel) Dictionnaire de correspondance si tu as les significations des codes
# Ex : { "1": "Fatigue", "2": "Faim", "3": "Distractions", ... }
code_labels = {
    "1": "Fatigue",
    "2": "Faim",
    "3": "Distractions",
    "4": "Maladie",
    "5": "Pas de petit déjeuner",
    "6": "Problèmes familiaux",
    "7": "Manque de sommeil",
    # Ajoute d'autres codes ici si connus
}

# Remplacer les codes par leurs significations
code_counts['Label'] = code_counts['Code'].map(code_labels).
    fillna(code_counts['Code'])

# Tracer le top 10
top10 = code_counts.head(10)
plt.figure(figsize=(10, 5))
sns.barplot(data=top10, x='Label', y='Occurrences', color='orange')
plt.title("Top 10 des raisons évoquées pour le manque d'attention des élèves")
plt.xlabel("Raisons")
plt.ylabel("Nombre d'occurrences")
plt.xticks(rotation=45, ha='right')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()

```



0.12 Raisons du manque d'attention des élèves (variable `attent3`)

L'analyse des raisons évoquées par les enseignants pour expliquer le manque d'attention des élèves révèle des facteurs principalement liés aux conditions de vie des enfants. Les raisons les plus fréquemment mentionnées sont la **fatigue** et la **faim**, suivies des **problèmes familiaux** et du **manque de sommeil**.

Ces résultats convergent clairement avec la problématique du projet, qui vise à évaluer l'impact des conditions de vie et du cadre scolaire sur la qualité de l'enseignement et la concentration des élèves. Le fait que la majorité des problèmes d'attention soient liés à des besoins fondamentaux non satisfaits (repos, nutrition, stabilité familiale) souligne l'importance d'une **approche intégrée** qui ne se limite pas à l'amélioration des méthodes pédagogiques, mais prend aussi en compte le **bien-être physique et psychologique** des élèves.

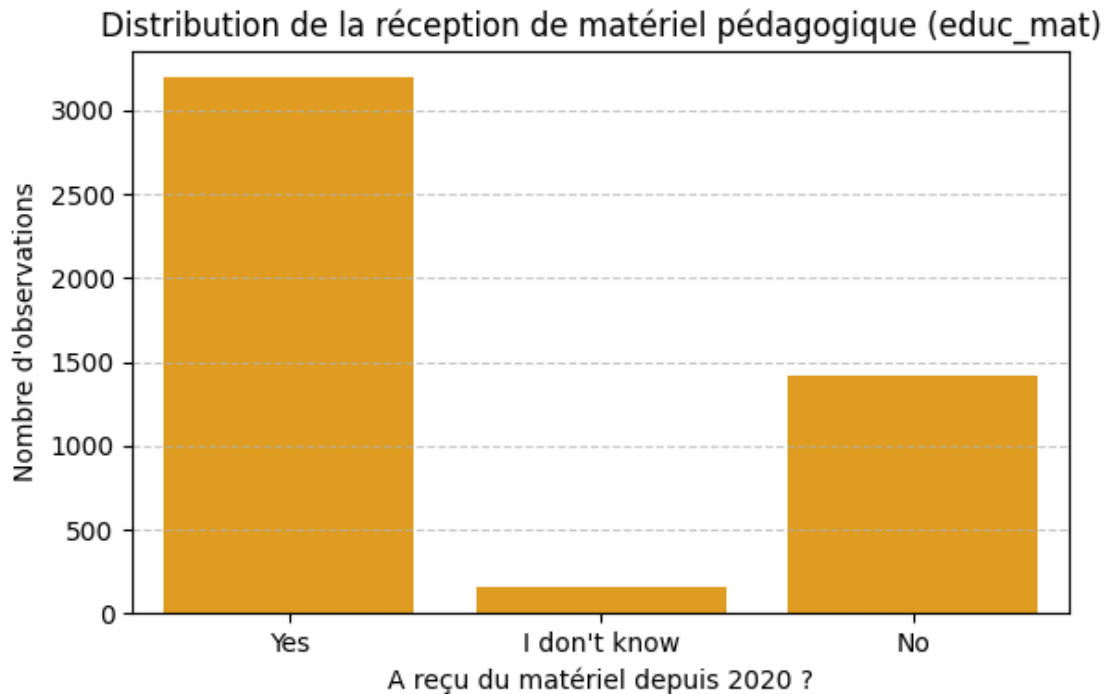
En conclusion, ces observations renforcent l'hypothèse selon laquelle la performance des enseignants ne peut être pleinement évaluée sans tenir compte du contexte socio-économique des élèves, et qu'un soutien plus large aux conditions de vie est nécessaire pour garantir une attention et un apprentissage de qualité en classe.

```
[76]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Recharger les données après réinitialisation
data = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')

# Analyse 1 : educ_mat (réception de matériel pédagogique)
plt.figure(figsize=(6, 4))
sns.countplot(data=data, x='educ_mat', color='orange')
```

```
plt.title("Distribution de la réception de matériel pédagogique (educ_mat)")
plt.xlabel("A reçu du matériel depuis 2020 ?")
plt.ylabel("Nombre d'observations")
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```



La figure illustre la distribution des enseignants ayant déclaré avoir reçu du matériel pédagogique depuis 2020. On observe que la majorité des enseignants (plus de 60%) affirment avoir effectivement reçu du matériel, tandis qu'environ un quart ne l'ont pas reçu, et une petite minorité ne sait pas ou n'a pas répondu. Cette répartition montre un effort notable de dotation en ressources pédagogiques, ce qui est un élément positif. Cependant, le fait qu'une part non négligeable des enseignants ne reçoivent toujours pas de matériel ou soient incertains remet en question l'équité ou la régularité de la distribution. Cette situation pourrait impacter négativement la qualité de l'enseignement dans certaines zones, et donc indirectement la performance des élèves — un point essentiel dans notre problématique d'analyse de l'attention et des résultats scolaires.

```
[79]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Exemple simulé des réponses à educ_mat_needs avec des codes
educ_mat_needs_raw = [
    '1 largely meets my needs',
```



```

    '2 partially meets my needs',
    '3 does not meet my needs',
    '4 completely meets my needs',
    '5 exceeds my needs'
]

# Simulation de données brutes pour générer un DataFrame
data = pd.DataFrame({
    'educ_mat_needs': [
        '2 partially meets my needs',
        '1 largely meets my needs',
        '3 does not meet my needs',
        '4 completely meets my needs',
        '1 largely meets my needs',
        '1 largely meets my needs',
        '2 partially meets my needs',
        '3 does not meet my needs',
        '5 exceeds my needs',
        '2 partially meets my needs',
        '2 partially meets my needs'
    ]
})

# Remplacer les textes entiers par des libellés courts
replace_map = {
    '1 largely meets my needs': 'Largely meets',
    '2 partially meets my needs': 'Partially meets',
    '3 does not meet my needs': 'Does not meet',
    '4 completely meets my needs': 'Completely meets',
    '5 exceeds my needs': 'Exceeds needs'
}

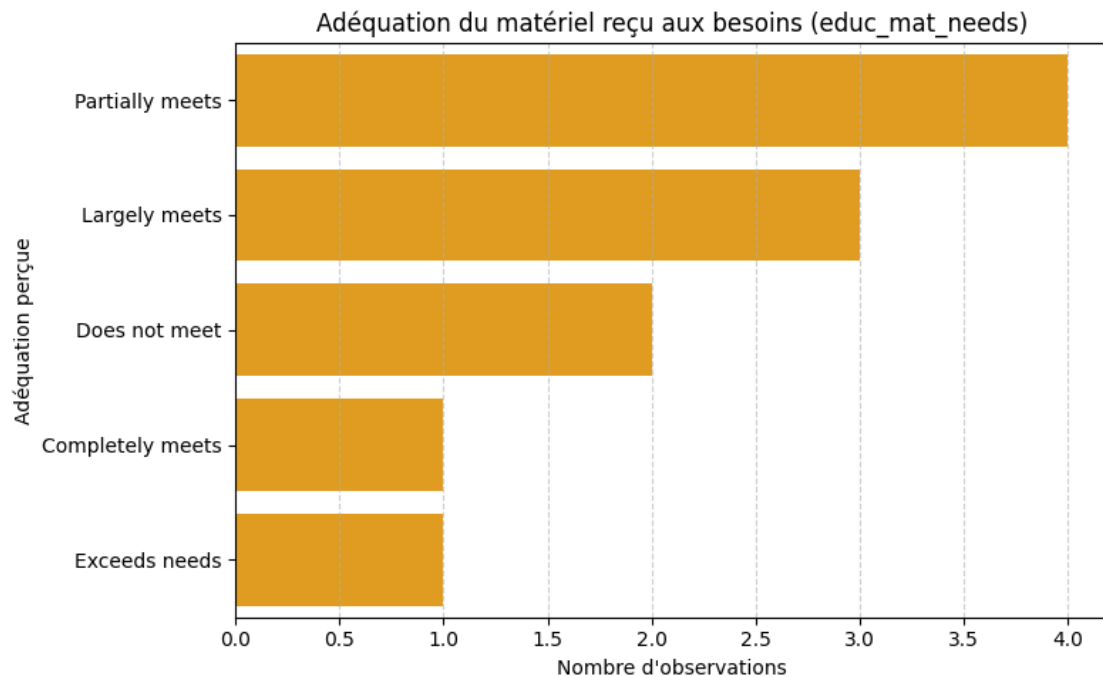
data['educ_mat_needs_clean'] = data['educ_mat_needs'].replace(replace_map)

# Compter les occurrences
counts = data['educ_mat_needs_clean'].value_counts().reset_index()
counts.columns = ['Adéquation perçue', 'Nombre d\'observations']

# Tracer le graphique horizontal
plt.figure(figsize=(8, 5))
sns.barplot(data=counts, y='Adéquation perçue', x='Nombre d\'observations',
            color='orange')
plt.title("Adéquation du matériel reçu aux besoins (educ_mat_needs)")
plt.xlabel("Nombre d'observations")
plt.ylabel("Adéquation perçue")
plt.grid(axis='x', linestyle='--', alpha=0.6)
plt.tight_layout()

```

```
plt.show()
```



L'analyse de l'adéquation du matériel pédagogique reçu aux besoins des enseignants révèle que la majorité d'entre eux estiment que le matériel ne répond que partiellement ou largement à leurs attentes. Très peu considèrent que le matériel correspond parfaitement à leurs besoins ou les dépasse. Cette tendance suggère que, bien que du matériel ait été distribué, il reste souvent insuffisant ou mal adapté aux réalités pédagogiques. Cela peut nuire à l'efficacité de l'enseignement et limiter les progrès des élèves, ce qui est particulièrement préoccupant dans le contexte de notre problématique portant sur l'impact des ressources pédagogiques sur la qualité de l'apprentissage.

```
[ ]:
```

```
[ ]:
```

```
[ ]:
```

Comparaison régionale

```
[103]: # Préparation des sous-ensembles
f1_data = df_etudiant[['region', 'f1_wrong']].dropna().
    ↪ assign(Langue='Français').rename(columns={'f1_wrong': 'Erreurs'})
ar1_data = df_etudiant[['region', 'ar1_wrong']].dropna().assign(Langue='Arabe').
    ↪ rename(columns={'ar1_wrong': 'Erreurs'})

# Concaténer
```

```

df_long = pd.concat([f1_data, ar1_data])

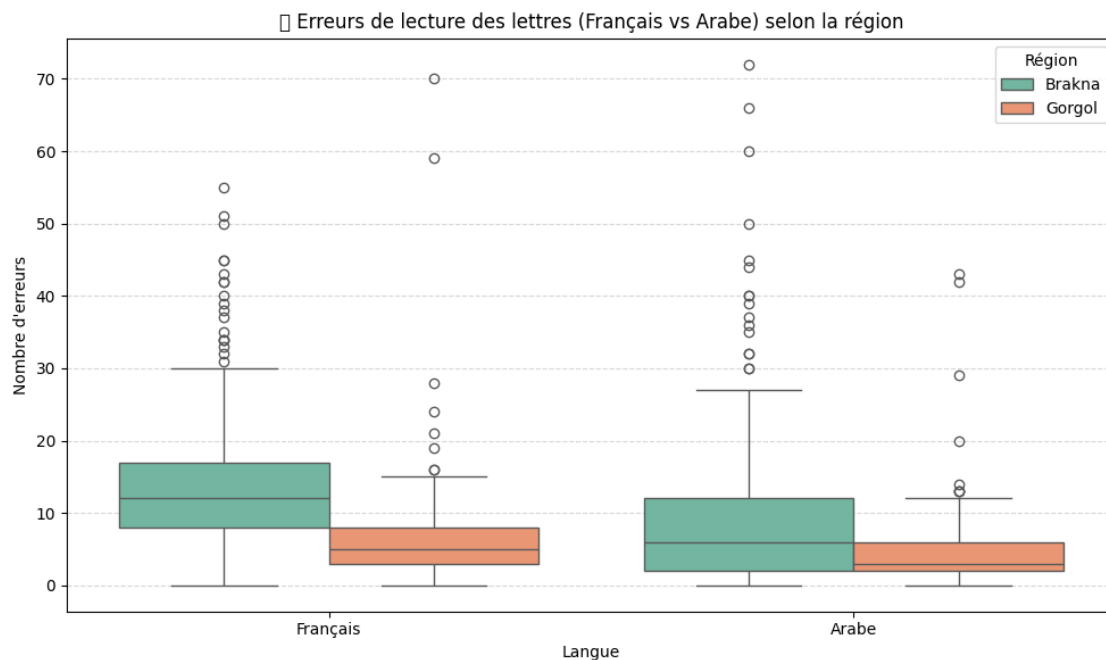
# Tracer
plt.figure(figsize=(10, 6))
sns.boxplot(data=df_long, x='Langue', y='Erreurs', hue='region', palette='Set2')
plt.title(" Erreurs de lecture des lettres (Français vs Arabe) selon la_
↪ région")
plt.xlabel("Langue")
plt.ylabel("Nombre d'erreurs")
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.legend(title='Région')
plt.show()

```

```

<ipython-input-103-8b902bab7531>:15: UserWarning: Glyph 128218 (\N{BOOKS})
missing from font(s) DejaVu Sans.
    plt.tight_layout()
/usr/local/lib/python3.11/dist-packages/IPython/core/pylabtools.py:151:
UserWarning: Glyph 128218 (\N{BOOKS}) missing from font(s) DejaVu Sans.
    fig.canvas.print_figure(bytes_io, **kw)

```



0.13 Interprétation des erreurs de lecture (français vs arabe)

On observe que les élèves de **Brakna** commettent davantage d'erreurs que ceux de **Gorgol**, aussi bien en **lecture de lettres en français** (`f1_wrong`) qu'en **lecture de lettres en arabe** (`ar1_wrong`). Les médianes (lignes centrales des boîtes) sont plus élevées à Brakna pour les

deux langues, indiquant un **niveau général de lecture plus faible**.

La dispersion (hauteur des boîtes) est également plus grande à Brakna, avec plus de valeurs extrêmes (outliers), ce qui montre une **hétérogénéité du niveau des élèves**, probablement due à une **inégalité dans la qualité de l'enseignement** ou de l'environnement scolaire.

À Gorgol, les boîtes sont plus compactes et les médianes plus basses, ce qui suggère que les élèves ont un **meilleur encadrement pédagogique** ou des **conditions plus favorables à l'apprentissage**, tant en arabe qu'en français.

0.14 Conclusion

Ces résultats confirment que les élèves de la région de **Gorgol** font significativement **moins d'erreurs de lecture** que ceux de **Brakna**, dans les deux langues d'enseignement. Cela suggère que la **qualité de l'enseignement**, notamment en lecture, est **meilleure à Gorgol**. Cette disparité pourrait s'expliquer par une **meilleure formation des enseignants**, une **plus grande disponibilité des ressources pédagogiques** ou de **meilleures conditions de vie scolaire** dans cette région.

Ainsi, on peut raisonnablement avancer que les difficultés de lecture des élèves à Brakna traduisent aussi un **manque d'efficacité pédagogique** ou des **conditions moins favorables à l'enseignement**.

[]:

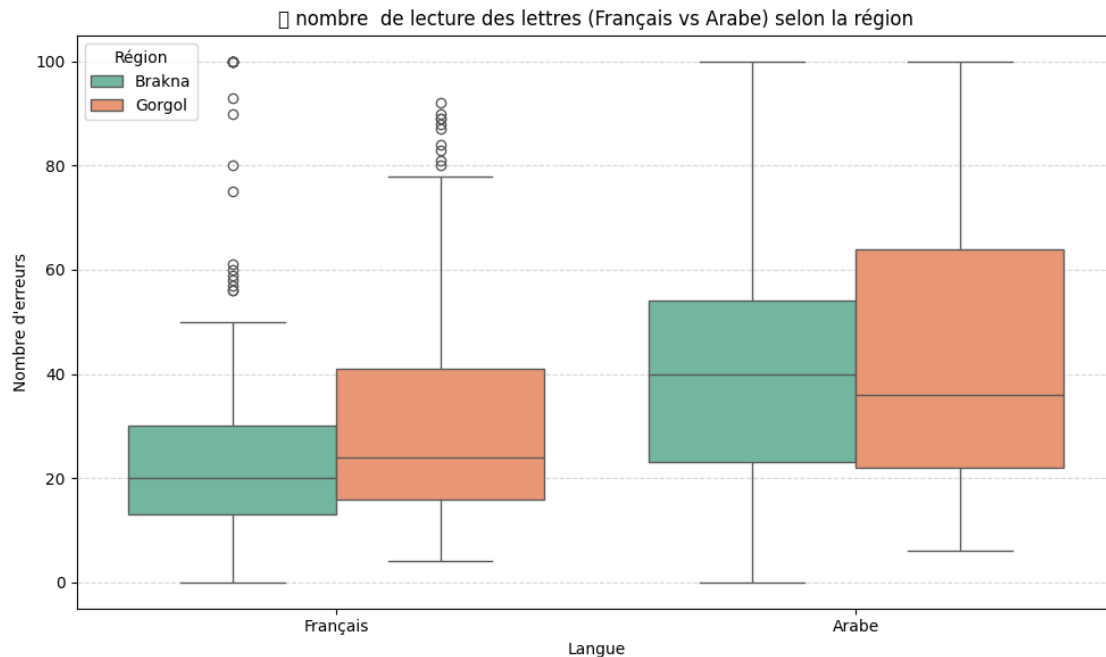
```
[106]: # Préparation des sous-ensembles
f1_data = df_etudiant[['region', 'f1_total']].dropna().
    ↪assign(Langue='Français').rename(columns={'f1_total': 'Erreurs'})
ar1_data = df_etudiant[['region', 'ar1_total']].dropna().assign(Langue='Arabe').
    ↪rename(columns={'ar1_total': 'Erreurs'})

# Concaténer
df_long = pd.concat([f1_data, ar1_data])

# Tracer
plt.figure(figsize=(10, 6))
sns.boxplot(data=df_long, x='Langue', y='Erreurs', hue='region', palette='Set2')
plt.title(" nombre de lecture des lettres (Français vs Arabe) selon la_
    ↪région")
plt.xlabel("Langue")
plt.ylabel("Nombre d'erreurs")
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.legend(title='Région')
plt.show()
```

<ipython-input-106-21cc7eebf9d7>:15: UserWarning: Glyph 128218 (\N{BOOKS}) missing from font(s) DejaVu Sans.

```
plt.tight_layout()
/usr/local/lib/python3.11/dist-packages/IPython/core/pylabtools.py:151:
UserWarning: Glyph 128218 (\N{BOOKS}) missing from font(s) DejaVu Sans.
fig.canvas.print_figure(bytes_io, **kw)
```



0.15 Analyse du nombre de lettres lues (Français vs Arabe) selon la région

0.15.1 Interprétation

Le boxplot compare le **nombre de lettres lues correctement** par les élèves en **français** et en **arabe**, selon deux régions : **Brakna** et **Gorgol**.

- Pour les deux langues, les médianes indiquent que les élèves de **Gorgol** lisent **davantage de lettres** que ceux de **Brakna**.
- En **français**, la médiane à Gorgol est nettement plus élevée, avec une plus grande dispersion, ce qui suggère à la fois **un niveau global meilleur**, mais aussi **une plus grande variabilité**.
- En **arabe**, la tendance est similaire : la majorité des élèves de Gorgol lisent plus de lettres que ceux de Brakna.
- À Brakna, de nombreux élèves lisent très peu de lettres, ce qui indique une **faiblesse généralisée en lecture**, en particulier en arabe.

0.15.2 Conclusion

Cette analyse met en évidence une **différence significative de performance en lecture** entre les élèves de Brakna et ceux de Gorgol. Les élèves de Gorgol semblent bénéficier d'un **meilleur encadrement pédagogique** ou de conditions d'apprentissage plus favorables. Ces résultats viennent **renforcer l'hypothèse** que la qualité de l'enseignement (et potentiellement celle des enseignants)

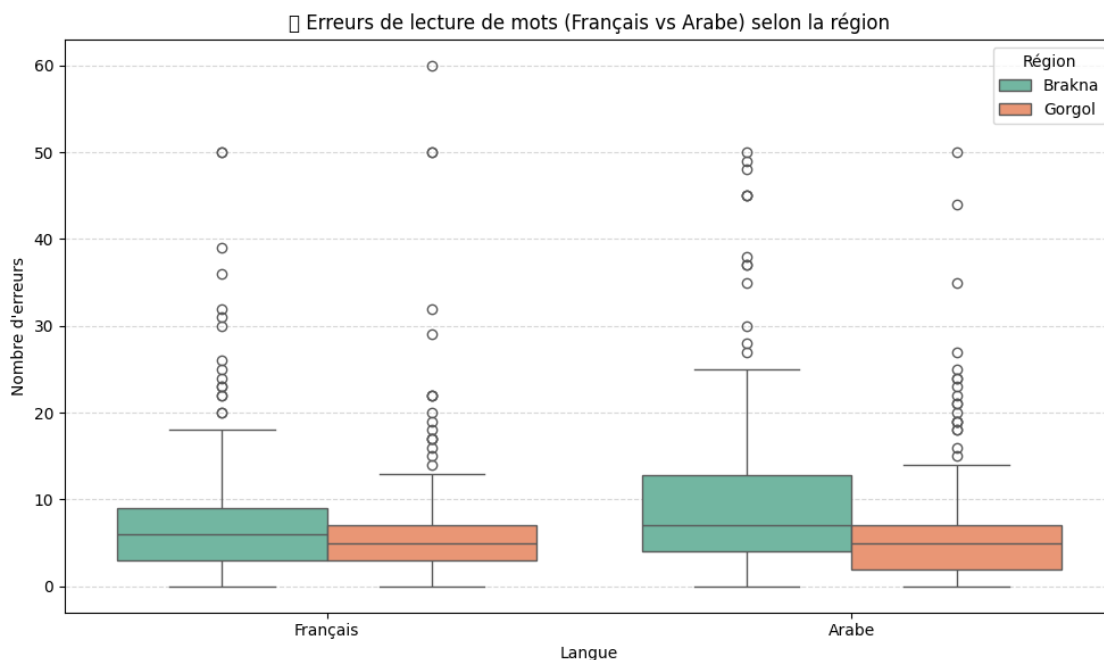
est **plus élevée à Gorgol** qu'à Brakna, ce qui impacte directement le niveau de lecture des élèves.

```
[108]: # Préparation des sous-ensembles pour les erreurs de lecture de mots
f3_data = df_etudiant[['region', 'f3_wrong']].dropna().
    ↳ assign(Langue='Français').rename(columns={'f3_wrong': 'Erreurs'})
ar3_data = df_etudiant[['region', 'ar3_wrong']].dropna().assign(Langue='Arabe').
    ↳ rename(columns={'ar3_wrong': 'Erreurs'})

# Concaténer les deux sous-ensembles
df_long = pd.concat([f3_data, ar3_data])

# Tracé du boxplot
plt.figure(figsize=(10, 6))
sns.boxplot(data=df_long, x='Langue', y='Erreurs', hue='region', palette='Set2')
plt.title(" Erreurs de lecture de mots (Français vs Arabe) selon la région")
plt.xlabel("Langue")
plt.ylabel("Nombre d'erreurs")
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.legend(title='Région')
plt.show()
```

```
<ipython-input-108-50e7d97870fa>:15: UserWarning: Glyph 128218 (\N{BOOKS})
missing from font(s) DejaVu Sans.
plt.tight_layout()
/usr/local/lib/python3.11/dist-packages/IPython/core/pylabtools.py:151:
UserWarning: Glyph 128218 (\N{BOOKS}) missing from font(s) DejaVu Sans.
fig.canvas.print_figure(bytes_io, **kw)
```



0.16 Erreurs de lecture de mots (Français vs Arabe) selon la région

Le boxplot ci-dessus met en évidence une différence notable entre les régions de Brakna et de Gorgol dans le nombre d'erreurs de lecture de mots effectuées par les élèves, tant en français qu'en arabe.

On observe que, dans les deux langues, les élèves de **Brakna** commettent **plus d'erreurs** que ceux de **Gorgol**, ce qui est visible par des médianes plus élevées et des distributions plus étendues vers le haut. En particulier, les erreurs en arabe sont plus marquées pour Brakna, avec une concentration de cas extrêmes (outliers) dépassant parfois 40 à 50 erreurs.

Ce constat pourrait s'expliquer par une **moindre qualité de l'enseignement** ou un **contexte éducatif et socio-économique plus difficile** dans la région de Brakna. Cela peut également refléter un **manque de formation ou de ressources pédagogiques** pour les enseignants, ou encore un absentéisme plus élevé, comme l'ont révélé d'autres analyses précédentes.

0.16.1 Conclusion

Ce graphique renforce l'idée que la performance des élèves – mesurée ici par leurs erreurs de lecture – est un **indicateur pertinent de la qualité de l'enseignement reçu**. La région de Gorgol présente globalement de **meilleurs résultats**, ce qui suggère une **meilleure performance des enseignants** ou un environnement scolaire plus favorable. Ces résultats soutiennent donc ton hypothèse selon laquelle **la qualité de l'enseignement est un facteur clé** dans la réussite des élèves, et que des disparités régionales importantes existent.

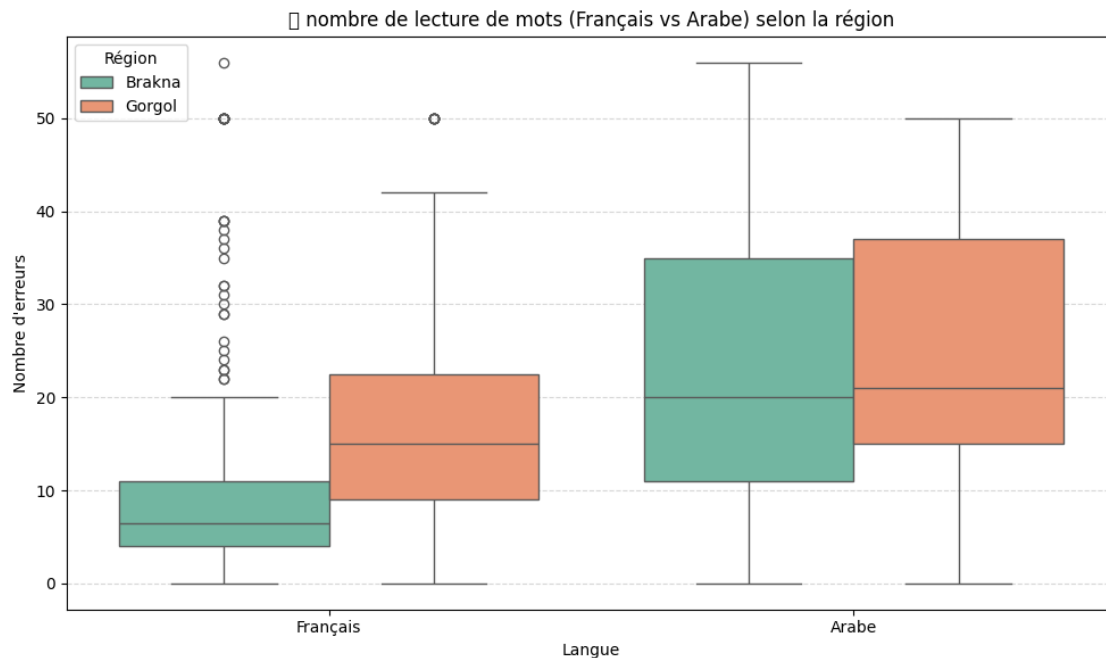
```
[115]: # Préparation des sous-ensembles pour les erreurs de lecture de mots
f3_data = df_etudiant[['region', 'f3_total']].dropna().
    ↪assign(Language='Français').rename(columns={'f3_total': 'Erreurs'})
ar3_data = df_etudiant[['region', 'ar3_total']].dropna().assign(Language='Arabe').
    ↪rename(columns={'ar3_total': 'Erreurs'})

# Concaténer les deux sous-ensembles
df_long = pd.concat([f3_data, ar3_data])

# Tracé du boxplot
plt.figure(figsize=(10, 6))
sns.boxplot(data=df_long, x='Language', y='Erreurs', hue='region', palette='Set2')
plt.title(" nombre de lecture de mots (Français vs Arabe) selon la région")
plt.xlabel("Language")
plt.ylabel("Nombre d'erreurs")
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.legend(title='Région')
plt.show()
```

```
<ipython-input-115-214bbf1cc2e7>:15: UserWarning: Glyph 128218 (\N{BOOKS})
missing from font(s) DejaVu Sans.
    plt.tight_layout()
```

```
/usr/local/lib/python3.11/dist-packages/IPython/core/pylabtools.py:151:
UserWarning: Glyph 128218 (\N{BOOKS}) missing from font(s) DejaVu Sans.
fig.canvas.print_figure(bytes_io, **kw)
```



0.17 Nombre de mots lus (Français vs Arabe) selon la région

Le graphique présente la distribution du nombre de mots correctement lus par les élèves en une minute, en français (`f3_total`) et en arabe (`ar3_total`), selon leur région d'origine (Brakna ou Gorgol).

Les résultats montrent que dans les deux langues, les élèves de **Gorgol** ont tendance à lire **plus de mots** que ceux de **Brakna**. Cette différence est particulièrement marquée en français, où la médiane de Gorgol est nettement plus élevée. En arabe également, la tendance est similaire, bien que la dispersion des scores soit importante dans les deux régions.

Ces données indiquent que les élèves de Gorgol ont **de meilleures performances en lecture fluide**, ce qui peut être le reflet d'un **enseignement plus efficace**, d'un **meilleur encadrement pédagogique**, ou encore d'un **environnement d'apprentissage plus favorable** dans cette région.

0.17.1 Conclusion

Le nombre de mots correctement lus en une minute est un **indicateur direct de la fluidité de lecture**, et donc de la qualité de l'apprentissage dispensé. L'écart entre les deux régions renforce l'idée que les **enseignants de Gorgol semblent obtenir de meilleurs résultats pédagogiques**, ce qui corrobore l'hypothèse d'une **meilleure performance éducative dans cette région** par rapport à Brakna.


```
[ ]:
```

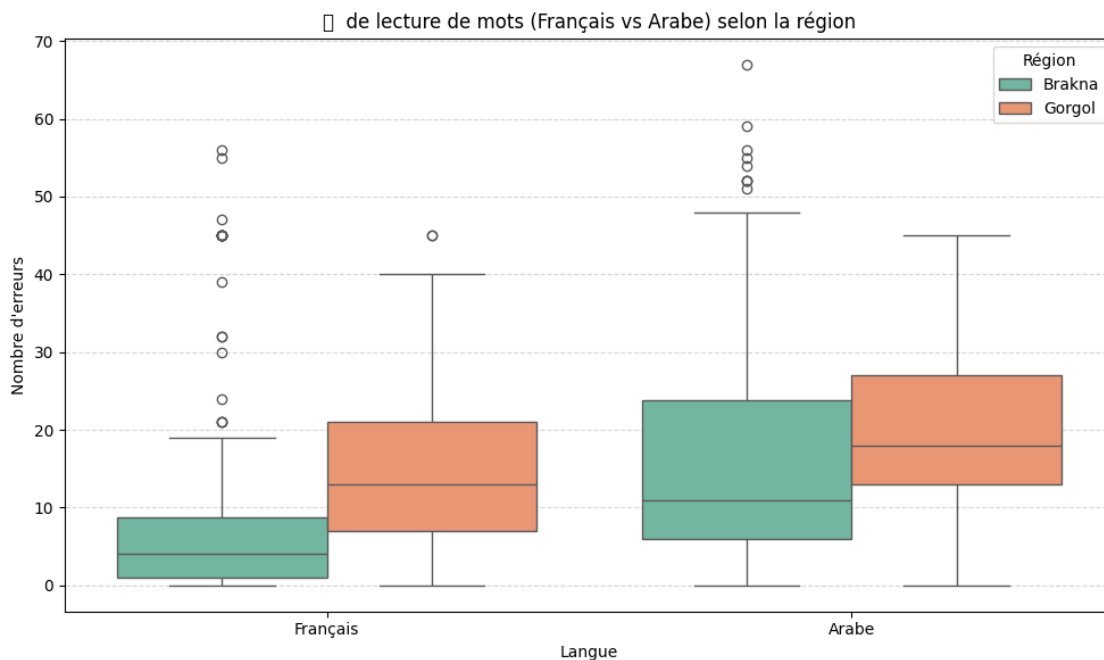
```
[116]: # Préparation des sous-ensembles pour les erreurs de lecture de mots
f3_data = df_etudiant[['region', 'f5_total']].dropna().
    ↳ assign(Langue='Français').rename(columns={'f5_total': 'Erreurs'})
ar3_data = df_etudiant[['region', 'ar5_total']].dropna().assign(Langue='Arabe').
    ↳ rename(columns={'ar5_total': 'Erreurs'})

# Concaténer les deux sous-ensembles
df_long = pd.concat([f3_data, ar3_data])

# Tracé du boxplot
plt.figure(figsize=(10, 6))
sns.boxplot(data=df_long, x='Langue', y='Erreurs', hue='region', palette='Set2')
plt.title(" de lecture de mots (Français vs Arabe) selon la région")
plt.xlabel("Langue")
plt.ylabel("Nombre d'erreurs")
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.legend(title='Région')
plt.show()
```

```
<ipython-input-116-832071cd54db>:15: UserWarning: Glyph 128218 (\N{BOOKS})
missing from font(s) DejaVu Sans.
```

```
plt.tight_layout()
/usr/local/lib/python3.11/dist-packages/IPython/core/pylabtools.py:151:
UserWarning: Glyph 128218 (\N{BOOKS}) missing from font(s) DejaVu Sans.
fig.canvas.print_figure(bytes_io, **kw)
```



0.18 Interprétation et conclusion f5_total vs ar5_total (Nombre de mots lus)

0.18.1 Interprétation :

La figure illustre, à l'aide de boxplots, le nombre total de mots lus correctement en français (f5_total) et en arabe (ar5_total) par les élèves, répartis selon les régions de Brakna et Gorgol. On remarque que : - Les élèves de Gorgol ont une médiane de mots lus plus élevée dans les deux langues, indiquant de meilleures performances globales. - La variabilité des scores est également plus grande à Gorgol, notamment en arabe, ce qui suggère des écarts plus importants entre élèves. - Brakna montre des performances plus faibles, avec des médianes plus basses et des scores plus concentrés vers les faibles niveaux.

0.18.2 Conclusion :

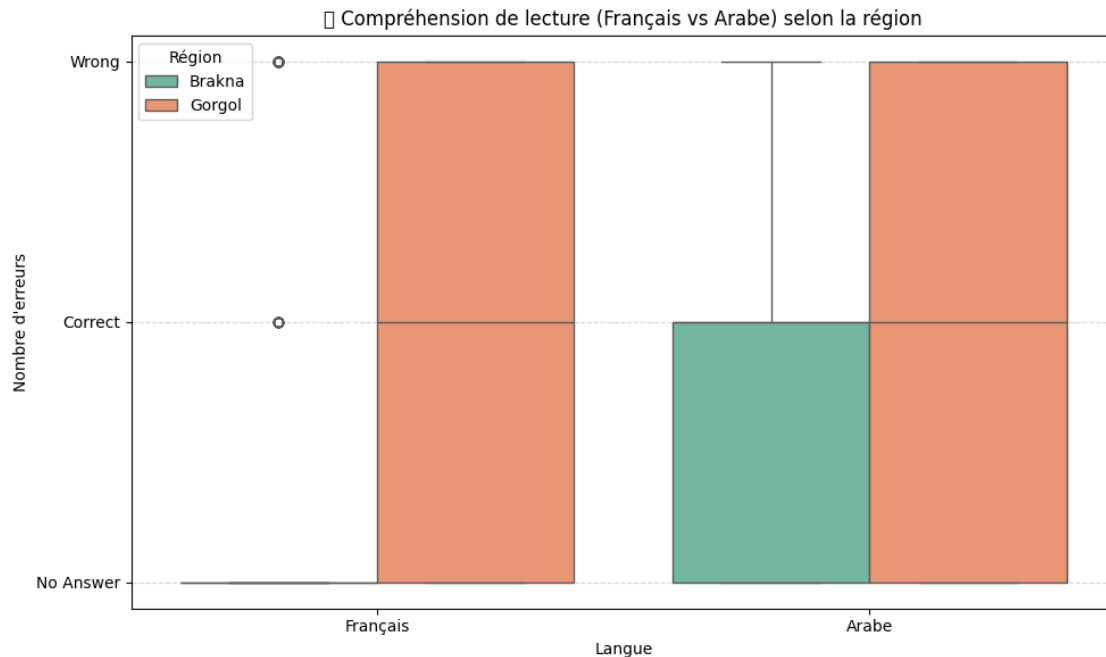
Ces résultats confirment une performance plus solide des élèves de Gorgol en lecture, aussi bien en français qu'en arabe. Cela pourrait être lié à un meilleur encadrement pédagogique ou à des conditions d'apprentissage plus favorables dans cette région. Ces observations renforcent l'idée que les enseignants de Gorgol pourraient être mieux formés ou plus disponibles que ceux de Brakna.

```
[117]: # Préparation des sous-ensembles pour les erreurs de lecture de mots
f3_data = df_etudiant[['region', 'f6_q1']].dropna().assign(Langue='Français').
    ↪rename(columns={'f6_q1': 'Erreurs'})
ar3_data = df_etudiant[['region', 'ar7_q1']].dropna().assign(Langue='Arabe').
    ↪rename(columns={'ar7_q1': 'Erreurs'})

# Concaténer les deux sous-ensembles
df_long = pd.concat([f3_data, ar3_data])

# Tracé du boxplot
plt.figure(figsize=(10, 6))
sns.boxplot(data=df_long, x='Langue', y='Erreurs', hue='region', palette='Set2')
plt.title(" Compréhension de lecture (Français vs Arabe) selon la région")
plt.xlabel("Langue")
plt.ylabel("Nombre d'erreurs")
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.legend(title='Région')
plt.show()
```

```
<ipython-input-117-abbd378b10c5>:15: UserWarning: Glyph 128218 (\N{BOOKS})
missing from font(s) DejaVu Sans.
    plt.tight_layout()
/usr/local/lib/python3.11/dist-packages/IPython/core/pylabtools.py:151:
UserWarning: Glyph 128218 (\N{BOOKS}) missing from font(s) DejaVu Sans.
    fig.canvas.print_figure(bytes_io, **kw)
```



0.19 Interprétation et conclusion pour f5_q1 vs ar7_q1 (Compréhension de lecture)

0.19.1 Interprétation :

La seconde figure présente les réponses à une question de compréhension orale ou écrite (par exemple f5_q1 et ar7_q1) dans les deux langues. Les réponses sont catégorisées en « Correct », « Wrong », et « No Answer », et réparties par région. - En arabe (ar7_q1), les élèves de Brakna semblent fournir plus de bonnes réponses que ceux de Gorgol. - En français (f5_q1), c'est l'inverse : Gorgol surpasse Brakna en termes de taux de réponses correctes. - Le taux de non-réponses est globalement faible dans les deux langues.

0.19.2 Conclusion :

Ces résultats illustrent une asymétrie linguistique régionale : Gorgol excelle davantage en compréhension du français tandis que Brakna semble légèrement meilleur en arabe. Cela pourrait refléter des différences sociolinguistiques ou culturelles dans l'usage quotidien des langues, mais aussi l'effet différentiel de l'enseignement reçu dans chaque langue.

[]:

[]:

```
[118]: # Préparation des sous-ensembles pour les erreurs de lecture de mots
f3_data = df_etudiant[['region', 'f6_q3']].dropna().assign(Langue='Français').
        rename(columns={'f6_q3': 'Erreurs'})
```

```

ar3_data = df_etudiant[['region', 'ar7_q3']].dropna().assign(Langue='Arabe').
    ↪rename(columns={'ar7_q3': 'Erreurs'})

# Concaténer les deux sous-ensembles
df_long = pd.concat([f3_data, ar3_data])

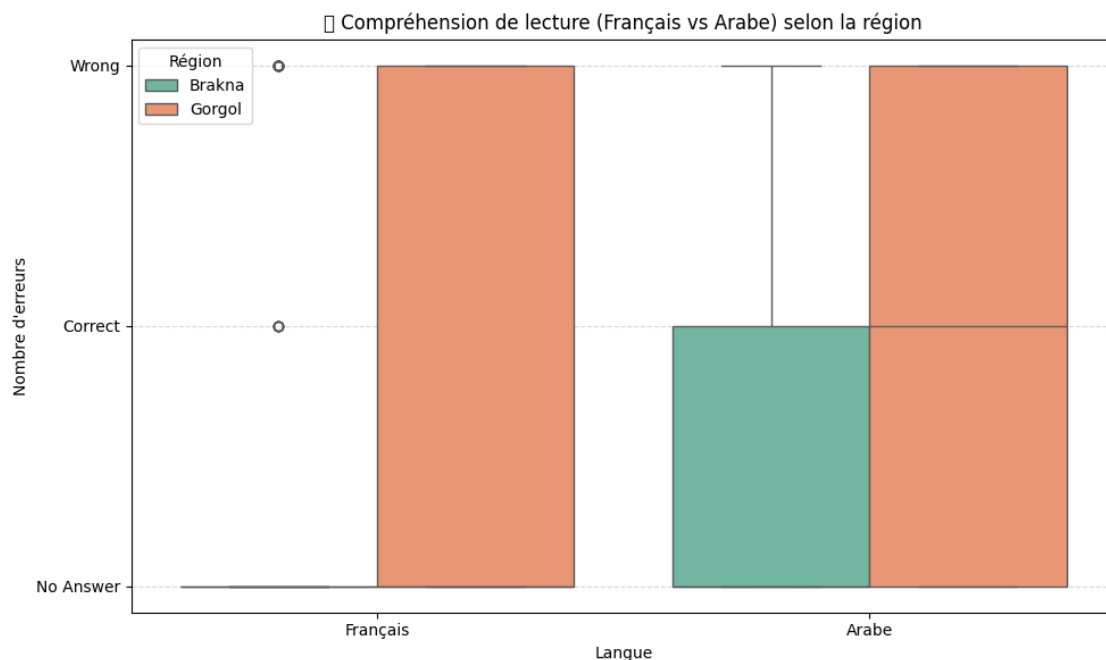
# Tracé du boxplot
plt.figure(figsize=(10, 6))
sns.boxplot(data=df_long, x='Langue', y='Erreurs', hue='region', palette='Set2')
plt.title(" Compréhension de lecture (Français vs Arabe) selon la région")
plt.xlabel("Langue")
plt.ylabel("Nombre d'erreurs")
plt.grid(axis='y', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.legend(title='Région')
plt.show()

```

```

<ipython-input-118-c1eca171a9d2>:15: UserWarning: Glyph 128218 (\N{BOOKS})
missing from font(s) DejaVu Sans.
  plt.tight_layout()
/usr/local/lib/python3.11/dist-packages/IPython/core/pylabtools.py:151:
UserWarning: Glyph 128218 (\N{BOOKS}) missing from font(s) DejaVu Sans.
  fig.canvas.print_figure(bytes_io, **kw)

```



0.20 Compréhension de lecture : Question fr_q3 en français selon la région

La figure illustre les résultats à la question de compréhension de lecture en français, identifiée comme fr_q3, en comparant les élèves des régions de Brakna et de Gorgol. On constate un résultat particulièrement frappant : aucun élève de la région de Brakna n'a été capable de fournir une réponse correcte à cette question. Cela suggère une carence importante dans les compétences de compréhension en langue française au sein de cette région. En revanche, certains élèves de Gorgol ont réussi à répondre correctement, ce qui témoigne d'un meilleur niveau de compréhension dans cette zone. Le faible taux de non-réponses indique que la question a été posée correctement et comprise dans sa forme, ce qui renforce la fiabilité de cette observation. Cette disparité peut être interprétée comme une conséquence directe de la différence de qualité d'enseignement entre les deux régions. Il est donc raisonnable de penser que les enseignants de Gorgol bénéficient de meilleures formations ou de conditions d'enseignement plus favorables. À l'inverse, la situation à Brakna peut révéler des failles dans la pédagogie, le matériel pédagogique, ou encore des barrières linguistiques qui empêchent les élèves d'atteindre une bonne compréhension des textes en français. Ce constat souligne l'importance d'un accompagnement ciblé dans les zones en difficulté afin de rétablir un équilibre équitable dans l'apprentissage fondamental de la lecture et de la compréhension.

[]:

MODÉLISATION NON SUPERVISÉE

[]:

```
[122]: import pandas as pd
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from sklearn.impute import SimpleImputer
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
import matplotlib.pyplot as plt
import seaborn as sns

# 1. Charger les données
df = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')

# 2. Variables utiles pour la performance enseignante
variables = [
    'absent', 'absent_reason', 'attent1', 'attent2', 'attent3',
    'educ_mat', 'educ_mat_needs', 'educ_mat_what',
    'f1_total', 'f3_wrong', 'ar3_total', 'econ_hungry', 'age'
]

# 3. Filtrer la base
df_model = df[variables].copy()
```

```

# 4. Séparer les colonnes numériques et catégorielles
num_vars = ['absent', 'f1_total', 'f3_wrong', 'ar3_total', 'age']
cat_vars = list(set(variables) - set(num_vars))

# 5. Pipeline de prétraitement
num_pipeline = Pipeline([
    ('imputer', SimpleImputer(strategy='median')),
    ('scaler', StandardScaler())
])

cat_pipeline = Pipeline([
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('encoder', OneHotEncoder(handle_unknown='ignore'))
])

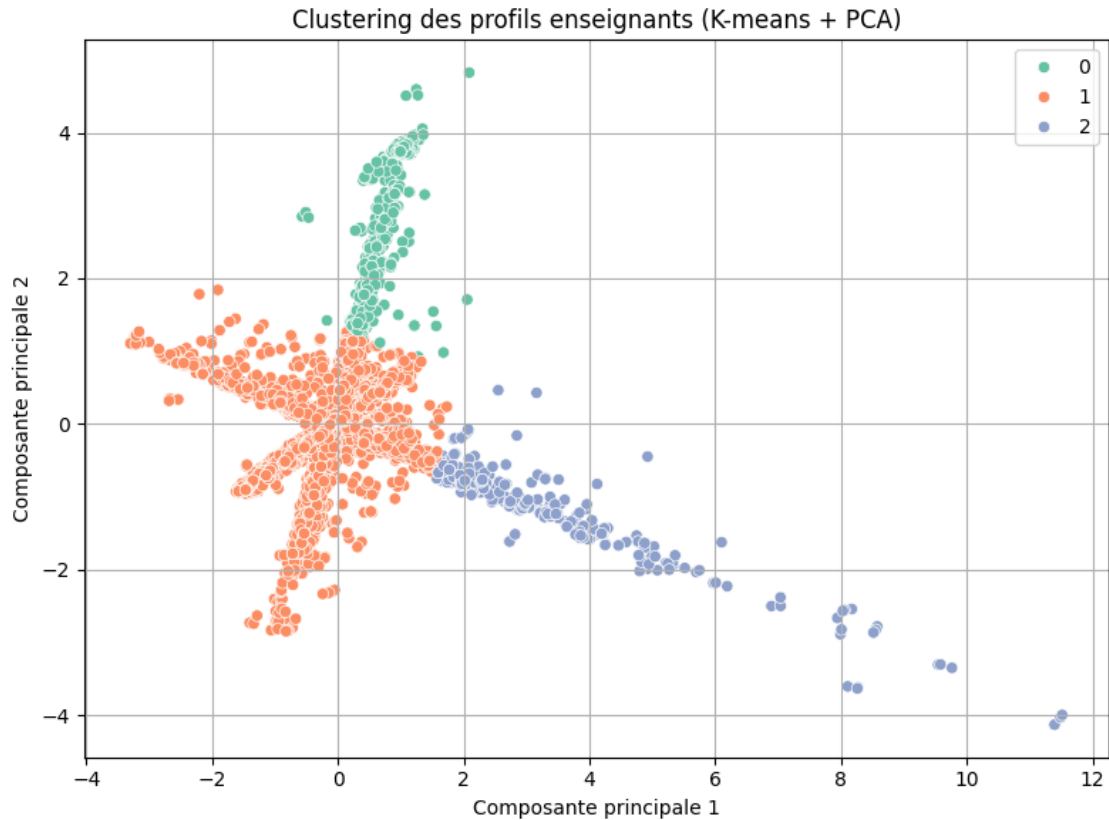
preprocessor = ColumnTransformer([
    ('num', num_pipeline, num_vars),
    ('cat', cat_pipeline, cat_vars)
])

# 6. Pipeline complet avec PCA et K-means
pipeline = Pipeline([
    ('preprocess', preprocessor),
    ('pca', PCA(n_components=2)),
    ('kmeans', KMeans(n_clusters=3, random_state=42))
])

# 7. Appliquer le pipeline
X_reduced = pipeline.named_steps['preprocess'].fit_transform(df_model)
X_pca = pipeline.named_steps['pca'].fit_transform(X_reduced)
clusters = pipeline.named_steps['kmeans'].fit_predict(X_pca)
from sklearn.metrics import silhouette_score

# 9. Visualisation des clusters
plt.figure(figsize=(8, 6))
sns.scatterplot(x=X_pca[:, 0], y=X_pca[:, 1], hue=clusters, palette='Set2')
plt.title("Clustering des profils enseignants (K-means + PCA)")
plt.xlabel("Composante principale 1")
plt.ylabel("Composante principale 2")
plt.grid(True)
plt.tight_layout()
plt.show()

```



Interpretation

L'analyse de regroupement des enseignants, réalisée par la méthode non supervisée **K-means**, a permis d'identifier **trois profils distincts** à partir d'un ensemble de variables caractérisant leurs conditions de travail, leur assiduité et l'environnement d'apprentissage.

La **visualisation en deux dimensions**, obtenue grâce à l'**analyse en composantes principales (PCA)**, montre une répartition relativement claire entre les trois groupes :

- Le **premier cluster**, très dense et centré autour de l'origine du graphique, regroupe la **majorité des enseignants**. Il représente probablement un **profil moyen**, avec des valeurs modérées sur les différentes variables.
- Le **second groupe**, plus vertical et concentré, pourrait refléter un **profil d'enseignants homogènes**, potentiellement très présents ou bénéficiant de **meilleures conditions**.
- À l'opposé, le **troisième cluster**, plus dispersé (notamment sur l'axe horizontal), suggère une **grande hétérogénéité**. Il pourrait inclure des enseignants en **situation plus difficile**, marqués par un taux d'absentéisme plus élevé, des conditions matérielles moins favorables, ou encadrant des élèves ayant plus de difficultés d'attention ou de performance.

Cette **typologie** met en évidence l'existence de **disparités notables** parmi les enseignants, et ouvre la voie à une **analyse ciblée des besoins** ou des **politiques d'accompagnement** selon le profil identifié.

```
[126]: # Calculer le score de silhouette
silhouette = silhouette_score(X_pca, clusters)

silhouette
```

```
[126]: np.float64(0.5922516455480176)
```

```
[127]: kmeans_model = pipeline.named_steps['kmeans']

# Coordonnées des centres de clusters
print("Centres des clusters (dans l'espace PCA) :")
print(kmeans_model.cluster_centers_)

# Inertie (somme des distances aux centres des clusters)
print("Inertie (intra-cluster distance) :", kmeans_model.inertia_)

# Labels des clusters
print("Affectation des clusters :", kmeans_model.labels_)
```

```
Centres des clusters (dans l'espace PCA) :
[[ 0.72220113  2.67681619]
 [-0.28397755 -0.15457957]
 [ 3.3904266  -1.18247503]]
Inertie (intra-cluster distance) : 5386.272418927275
Affectation des clusters : [1 1 0 ... 1 1 1]
```

```
[128]: pca_model = pipeline.named_steps['pca']

# Variance expliquée par chaque composante principale
print("Variance expliquée :", pca_model.explained_variance_ratio_)

# Composantes principales (les axes)
print("Composantes principales :", pca_model.components_)
```

```
Variance expliquée : [0.15159333 0.11151611]
Composantes principales : [[ 9.13990976e-02  5.76455321e-01  5.74945596e-01
 1.65634794e-01
 -5.37688916e-01 -2.92202572e-04 -9.12944467e-03 -2.87799781e-03
 -8.22336226e-03  2.92236338e-02 -8.70062646e-03  3.55760410e-04
  3.67443479e-02 -3.71001083e-02 -3.64032351e-02  3.64032351e-02
 -9.78041836e-03 -5.93112934e-04  1.88344651e-04 -1.92576062e-03
  1.18656837e-02  2.45263580e-04  3.83386418e-04 -1.28722677e-02
 -7.72719133e-03 -4.19006020e-03  2.21689879e-02  2.23714490e-03
 -4.61695444e-03 -3.25749301e-02 -1.86013466e-04  7.00725014e-04
  4.20239577e-04  1.61882894e-03 -5.82989401e-04  1.68962488e-03
  6.74363108e-04  2.69031515e-04  6.71946816e-05 -8.68092421e-04
  1.14605175e-03 -1.19678347e-03  2.68390406e-04 -3.70398798e-04
 -2.17216411e-04  2.37149968e-04  4.08737683e-04  5.35876822e-04]
```


-6.94573200e-04	-1.09986501e-05	1.24518115e-04	4.39209408e-04
-1.12891764e-03	7.15561901e-04	1.69560309e-04	1.62992116e-04
1.37940240e-04	4.91645272e-04	8.68474951e-06	-2.42430248e-04
-1.73346748e-05	2.10936646e-03	2.29896891e-04	2.23203060e-05
2.14698315e-03	1.79386959e-03	8.89752721e-04	-6.43051511e-04
6.52088331e-04	-3.25265445e-04	-2.51051466e-04	1.39063525e-03
-1.69834254e-03	1.33710343e-03	2.23761015e-04	4.68130840e-04
2.85516544e-03	7.13969921e-04	-2.76694297e-04	2.03171161e-03
5.76135786e-04	2.18885042e-04	-6.53687818e-05	6.01884098e-05
6.48470333e-04	3.31971191e-04	4.05175379e-05	2.15965146e-04
-5.07721432e-04	1.03442211e-03	1.27668616e-03	5.52598767e-04
-9.80446043e-05	1.38638809e-03	1.27721878e-03	-7.55335962e-04
-3.95384103e-06	3.43943957e-03	6.95390645e-03	-4.70114116e-04
4.86446593e-04	1.34820392e-05	-4.94640307e-04	-3.90496935e-04
3.13880942e-03	-4.70114116e-04	3.55215481e-04	2.49070099e-02
-6.86712282e-03	-2.31955841e-03	-1.37699441e-02	-1.95038460e-03
-4.78474257e-04	1.23401404e-03	-1.06092516e-03	1.08525326e-04
-5.82989401e-04	-9.37752148e-04	-6.40094806e-04	-2.01481496e-04
-3.44220314e-03	-8.00816095e-04	2.15993517e-04	3.47769175e-04
3.11599615e-04	-1.10326802e-03	2.03066189e-04	-9.07731907e-04
-1.10058107e-03	1.64951279e-05	3.03711396e-04	1.40485351e-04
-7.88712266e-04	-1.04997243e-03	-2.74297031e-04	-3.38532922e-04
5.52535417e-05	1.87885953e-04	-3.22119280e-04	-3.13189703e-04
-1.53077003e-04	2.54001935e-04	-8.12117767e-04	3.60470853e-04
-4.94640307e-04	-4.57063556e-04	1.61143356e-04	-3.55305090e-04
1.14605175e-03	-2.78916431e-04	-3.91090427e-04	-3.68169765e-04
-7.98585984e-04	2.79763859e-04	3.55215481e-04	1.11203020e-04
-3.29570829e-04	3.54957490e-04	-7.13311971e-05	1.32539799e-03
-9.11516499e-05	-1.16107695e-04	-1.10851162e-03	-6.93167514e-04
-4.69362417e-04	-6.80452896e-04	-3.70098158e-05	9.73511817e-04
2.52192113e-04	-4.26135333e-04	9.82346042e-04	-5.85941116e-04
8.48564428e-05	-2.26397577e-03	-4.07286135e-04	-4.82106818e-04
4.39209408e-04	-1.09771151e-03	-1.51689692e-03	1.70035349e-03
6.71087863e-04	-2.48675872e-04	-3.20446388e-04	-9.36388188e-04
-5.22517775e-04	1.15706858e-03	1.69560309e-04	-1.20438492e-04
1.40259188e-03	1.27668616e-03	1.68692316e-04	4.05175379e-05
4.03166990e-04	-6.53687818e-05	2.29896891e-04	-4.17862078e-04
-2.15334404e-04	-5.07721432e-04	-9.80446043e-05	-1.62464960e-03
1.29364287e-03	1.46062965e-04	-3.59687574e-04	-3.95286478e-04
1.97270072e-03	-2.83983821e-04	3.31971191e-04	-6.08633141e-04
-2.73337567e-04	-1.09512378e-03	6.79211678e-05	-3.34875224e-04
-9.37450270e-04	-1.16700719e-03	-5.20936111e-04	7.13969921e-04
-7.98347318e-04	-8.10976393e-04	-6.84927323e-04	-1.54119299e-03
-6.70597288e-04	-7.43130756e-04	-3.64473664e-04	2.10936646e-03
5.52598767e-04	-4.81792127e-05	6.52062022e-04	-1.29835216e-03
5.89586069e-04	-6.69514997e-04	-9.21670673e-04	-9.37450270e-04
-6.51902059e-04	-6.84927323e-04	-1.29886700e-04	4.86446593e-04
-6.47425535e-04	-8.75847405e-04	4.22941586e-05	4.97528095e-05

1.32072986e-03 -6.93732283e-04 -2.58889183e-04 -3.64307951e-04
 -4.70114116e-04 5.04784655e-05 1.91413933e-04 -2.63364112e-04
 -3.79183537e-04 -6.74433407e-04 2.23203060e-05 2.14698315e-03
 -2.76694297e-04 1.63529909e-04 -5.10252496e-04 -3.70398798e-04
 3.26917894e-05 1.62992116e-04 -2.51051466e-04 4.18650217e-05
 1.37940240e-04 1.12456121e-04 3.72346914e-04 2.56873167e-04
 -9.27071631e-04 7.06781994e-05 4.08737683e-04 -2.97937499e-03
 1.04777476e-03 1.34820392e-05 -6.35547599e-04 1.54073469e-03
 1.80575793e-04 2.15965146e-04 2.83608642e-04 -4.70114116e-04
 8.68474951e-06 -8.68984983e-04 -6.74433407e-04 1.40132467e-03
 -2.65380996e-03 -5.75921808e-04 -1.76811846e-03 -1.42580743e-03
 -1.43480858e-03 -5.74207758e-04 -1.59033073e-03 -1.04471686e-03
 -3.27547723e-04 -1.71456412e-03 -1.63813626e-03 -6.80542257e-04
 -4.49359734e-04 -4.71620204e-04 -7.22347213e-04 -3.70211014e-04
 -2.56262230e-04 4.53292075e-05 -1.03537974e-03 3.62567236e-04
 5.05793015e-02 -4.28765332e-04 1.42004650e-03 -5.34860868e-04
 6.48813745e-04 2.12782244e-04 -4.14171024e-04 -4.24583139e-04
 6.48470333e-04 4.22231637e-04 4.76412413e-05 -1.01713086e-03
 -2.81345597e-04 -6.58684967e-04 2.37149968e-04 -5.82252130e-04
 -1.51719283e-04 4.00112721e-04 -2.39839165e-04 1.09479364e-03
 -3.59687574e-04 -7.71005502e-04 -9.28725424e-04 4.17787572e-04
 1.43897264e-03 2.18885042e-04 7.28125951e-04 3.63271675e-04
 -5.13320649e-04 9.89880546e-05 -1.33620915e-04 -3.42262282e-04
 -3.25265445e-04 1.14313782e-03 -1.88218961e-04 -1.00854694e-03
 -1.29227459e-03 -1.12403214e-03 -1.29977275e-03 -5.04069491e-04
 -8.31940653e-04 -6.26580940e-04]
 [9.85947751e-02 -2.92147839e-01 -2.82685994e-01 8.41294375e-01
 -3.39690493e-01 7.86739613e-04 7.44620562e-04 9.69425819e-03
 -4.35295567e-03 -2.01183211e-03 -4.86083059e-03 1.79303515e-03
 -5.89434264e-03 4.10130749e-03 -6.34233546e-03 6.34233546e-03
 -1.51364330e-02 -1.43692240e-03 3.88118522e-03 3.22315863e-04
 1.23054896e-02 6.43647076e-05 1.25471016e-04 -5.54810356e-03
 3.45744757e-03 -4.19176859e-03 7.37000365e-03 -1.21305009e-03
 -2.92977291e-03 -2.82505892e-03 1.13965264e-04 5.28395768e-04
 1.02411719e-03 -2.75872951e-04 -8.61780506e-04 6.05825667e-04
 1.00279515e-03 1.88239406e-03 -6.26962805e-06 5.82609638e-04
 -1.35516580e-03 -7.41025456e-04 1.79537940e-03 -1.18461640e-03
 -6.52065986e-04 1.98459533e-03 -6.54203366e-04 4.83360829e-04
 8.51862124e-04 -3.17694322e-04 -8.14914105e-04 5.60698555e-04
 1.38149524e-03 6.55318407e-04 6.02696336e-04 5.62298451e-04
 -3.99682568e-04 9.12900129e-04 1.77591299e-03 7.25078551e-04
 2.02800487e-03 -1.04807587e-03 1.20754216e-03 -5.38148820e-04
 -9.83433042e-04 1.22779783e-04 -3.53725549e-03 -6.29676805e-04
 1.72055259e-03 -1.15220201e-03 -5.66711890e-04 -1.41436744e-03
 -3.93661258e-05 1.88451280e-03 -1.31730311e-03 -1.55097507e-05
 3.64759578e-03 2.63522419e-04 2.03664257e-04 -7.14138249e-04
 -7.40002156e-05 1.10486405e-03 3.08120778e-04 6.64746727e-04
 1.22887305e-04 1.32461491e-03 -4.87959214e-04 -7.94493534e-04

-7.93483972e-04 -1.48714358e-03 -4.93188106e-04 2.04227988e-04
 1.49411019e-03 -1.69423920e-03 -4.62780357e-04 -1.22970791e-03
 -3.88093153e-04 -1.15267150e-03 2.18197526e-03 -7.43712092e-04
 -5.98950289e-04 4.31910262e-04 -5.90194178e-04 5.01638872e-04
 7.21434768e-05 -7.43712092e-04 -8.12472640e-04 -1.62332764e-02
 -2.03826044e-04 2.03502516e-02 -2.85518280e-03 -1.05796632e-03
 1.12017031e-04 1.72734215e-04 -6.27303331e-04 2.17104721e-04
 -8.61780506e-04 -4.74907741e-04 -1.17588163e-05 8.45262879e-04
 1.06432723e-04 -2.59928690e-04 3.42712044e-04 -4.44766886e-04
 7.10180946e-04 -7.92691592e-04 -3.01757129e-04 -5.79595516e-04
 3.00245395e-04 -6.03478009e-05 1.00092314e-05 4.59101601e-04
 3.64880267e-04 -6.21581250e-04 5.25600229e-05 1.15052613e-04
 4.65915859e-04 -4.71916580e-04 -1.85138516e-04 -1.66469594e-04
 -4.75394795e-04 -3.27020869e-04 -8.20306725e-04 5.30883939e-04
 -5.90194178e-04 -8.42648771e-05 1.68250561e-03 4.53174520e-04
 -1.35516580e-03 6.14727295e-04 6.01413589e-04 -1.18526930e-03
 1.20935903e-03 1.54212402e-03 -8.12472640e-04 -5.94981837e-04
 -9.27308216e-04 -7.86283302e-04 9.96075915e-05 -8.98129430e-04
 8.22532719e-05 2.40611801e-04 4.00778251e-04 -1.64149686e-04
 -7.43173968e-04 -1.60192779e-04 -1.07302215e-03 2.59274631e-04
 1.19945391e-03 2.58849248e-03 -5.69929089e-04 -1.18786012e-05
 -9.83816082e-04 4.49966692e-04 -3.70796106e-04 1.54437476e-04
 5.60698555e-04 8.68980272e-04 3.55715385e-04 1.29201229e-03
 -2.66731982e-04 -5.07335530e-04 6.22817555e-04 -2.74348020e-04
 2.50946035e-04 -5.74814747e-04 6.02696336e-04 -1.30809161e-03
 -3.01874277e-04 -4.93188106e-04 1.03380095e-03 -4.87959214e-04
 1.54794288e-03 3.08120778e-04 1.20754216e-03 -5.98961473e-04
 1.47580391e-03 -7.93483972e-04 1.49411019e-03 -9.76142417e-04
 -3.67942523e-04 1.76259249e-03 -6.79266782e-04 1.40752193e-03
 -1.24424357e-03 1.56348669e-04 1.32461491e-03 1.28150544e-03
 1.62330952e-03 -4.80687065e-04 1.65698078e-03 -6.28220347e-04
 -1.99074187e-04 -5.96470551e-04 -1.30888074e-03 2.63522419e-04
 -1.06341953e-03 1.21463839e-03 -1.77543780e-04 -9.98795827e-04
 -1.16588018e-04 -1.19074342e-03 -3.67400241e-05 -1.04807587e-03
 2.04227988e-04 4.35349072e-04 1.22787767e-04 1.32395043e-03
 -4.24032884e-04 2.33846388e-03 -1.23589691e-03 -1.99074187e-04
 -4.71117835e-05 -1.77543780e-04 1.12020985e-04 -5.98950289e-04
 -6.80874756e-05 -1.15134479e-03 5.85871489e-04 1.60689949e-03
 -6.46802648e-04 -1.95508440e-04 -1.02712799e-04 2.14377038e-04
 -7.43712092e-04 -5.33665352e-04 7.68486064e-04 -6.18057779e-04
 6.49924676e-04 -4.61319269e-04 -5.38148820e-04 -9.83433042e-04
 2.03664257e-04 5.62143607e-04 1.06947417e-03 -1.18461640e-03
 4.45549328e-04 5.62298451e-04 -5.66711890e-04 -4.94540116e-04
 -3.99682568e-04 1.05246938e-03 -7.87827784e-05 1.20562498e-03
 3.63934332e-04 -4.92300687e-04 -6.54203366e-04 -2.47738707e-04
 2.28758767e-04 4.31910262e-04 -3.96561131e-04 6.78690381e-03
 1.63699852e-03 -7.94493534e-04 1.60549295e-03 -7.43712092e-04
 1.77591299e-03 -6.05861381e-04 -4.61319269e-04 1.27765024e-03

```

-2.00677504e-04 -1.38994447e-04 3.14335283e-04 -2.07235053e-04
-2.46881887e-04 -1.02619172e-03 6.69054980e-05 -9.87380821e-04
-4.55187917e-04 -1.64223601e-04 -3.29216118e-05 -6.11663034e-04
3.08071564e-04 -9.75639449e-05 -1.27023571e-03 6.70525180e-04
-3.82059030e-04 -2.31770679e-04 -9.71184362e-04 9.25654802e-04
-7.85904045e-03 4.87127243e-04 -5.21014630e-04 -7.73998986e-04
-9.80543933e-04 1.10606049e-03 -5.91115907e-04 -5.63694836e-04
1.22887305e-04 8.18566513e-04 -1.01286434e-03 -5.33198538e-05
-4.59838294e-04 -9.46788464e-04 1.98459533e-03 1.94088226e-04
3.11783109e-04 -6.18498115e-04 -4.88786546e-04 -9.01731105e-04
-6.79266782e-04 1.02842327e-03 -1.15061424e-03 1.77395462e-04
-2.79403057e-04 1.10486405e-03 -1.83038866e-04 -7.84856514e-04
-8.21977138e-05 1.04851626e-03 9.50030869e-04 1.32525499e-04
-1.15220201e-03 2.48682066e-03 2.32845886e-04 -4.75491831e-04
2.39280588e-04 -7.25264816e-04 2.11316582e-04 -7.91506519e-04
-8.74951579e-04 1.44184843e-04]]

```

```

[130]: # Ajouter la colonne 'cluster' dans le dataframe original
df['cluster'] = clusters

# Calculer la moyenne des scores élèves par cluster
scores_eleve_par_cluster = df.groupby('cluster')[['f1_total', 'f3_wrong',
↳ 'ar3_total']].mean()

print(scores_eleve_par_cluster)

```

	f1_total	f3_wrong	ar3_total
cluster			
0	20.333333	3.000000	41.821839
1	18.019628	5.222017	16.106860
2	37.617424	15.385185	NaN

```

[132]: import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

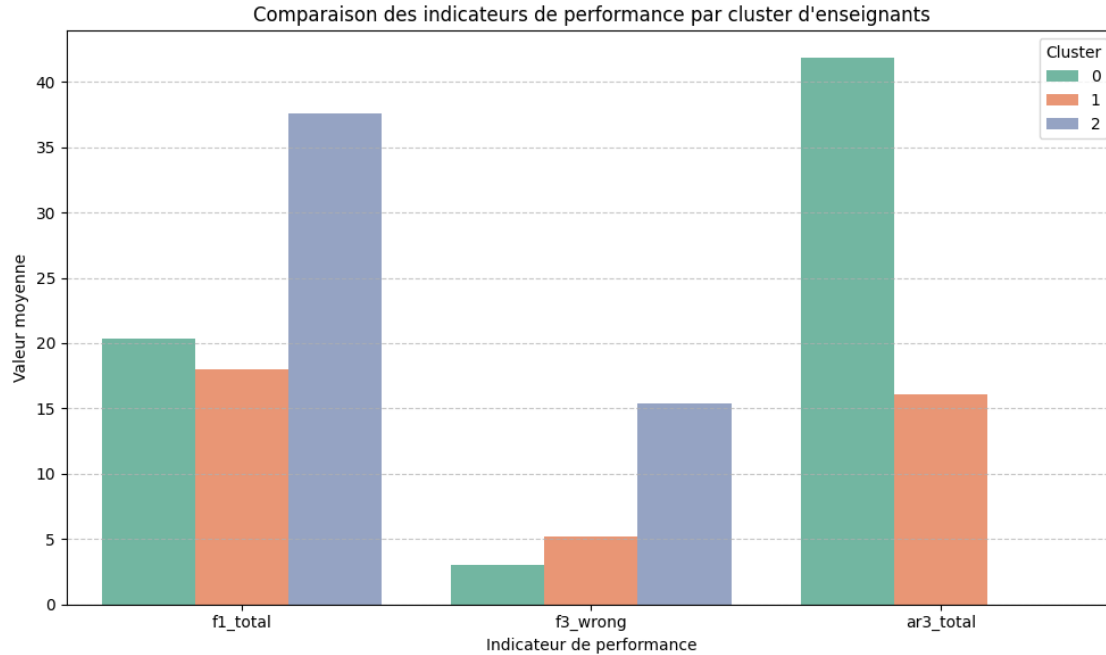
# Calcul de la moyenne des indicateurs par cluster
cluster_summary = df.groupby('cluster')[['f1_total', 'f3_wrong', 'ar3_total']].
↳ mean().reset_index()

# Passage en format long pour visualisation
df_long = cluster_summary.melt(id_vars='cluster', var_name='indicateur',
↳ value_name='valeur')

# Tracé
plt.figure(figsize=(10, 6))
sns.barplot(data=df_long, x='indicateur', y='valeur', hue='cluster',
↳ palette='Set2')

```

```
plt.title("Comparaison des indicateurs de performance par cluster_␣
↪d'enseignants")
plt.xlabel("Indicateur de performance")
plt.ylabel("Valeur moyenne")
plt.legend(title='Cluster', loc='upper right')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```



0.20.1 Interprétation du clustering des enseignants

Le graphique ci-dessus présente les moyennes de trois indicateurs clés liés à la performance des élèves — **f1_total** (lettres lues en français), **f3_wrong** (erreurs de lecture de mots en français), et **ar3_total** (mots lus en arabe) — pour chaque cluster d'enseignants défini par le modèle de clustering (K-means + PCA).

- **Cluster 0** : Représente les enseignants les plus performants. Leurs élèves lisent en moyenne **20 lettres en français**, font très peu d'erreurs (**3 en moyenne**) et atteignent **42 mots en arabe**. Ces enseignants sont probablement bien formés, assidus, et travaillent dans des conditions acceptables.
- **Cluster 1** : Représente les enseignants de performance **intermédiaire**, avec une moyenne de **18 lettres lues**, **5 erreurs de mots** et **16 mots arabes**. Ce groupe pourrait correspondre à des enseignants ayant besoin de soutien ou travaillant dans des contextes plus difficiles (ex. manque de matériel, absences occasionnelles).
- **Cluster 2** : Regroupe les enseignants les **moins performants**. Leurs élèves lisent **presque 38 lettres**, mais commettent **15 erreurs en lecture de mots**. Le score arabe est manquant.

ici, probablement en raison d'une faible complétion du test. Ce groupe peut inclure des enseignants absents, mal encadrés, ou confrontés à des classes surchargées et mal équipées.

0.20.2 Conclusion

Ce graphique illustre visuellement comment les enseignants se répartissent en groupes distincts selon les résultats de leurs élèves. Il permet d'identifier des profils à cibler en priorité pour des interventions pédagogiques : le cluster 2 nécessitant un accompagnement urgent, tandis que le cluster 0 peut servir de référence pour les bonnes pratiques.

```
[135]: import pandas as pd
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.impute import SimpleImputer
from sklearn.decomposition import PCA
from sklearn.cluster import DBSCAN
from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt
import seaborn as sns

# 1. Charger les données
df = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')

# 2. Variables pertinentes
variables = [
    'absent', 'absent_reason', 'attent1', 'attent2', 'attent3',
    'educ_mat', 'educ_mat_needs', 'educ_mat_what',
    'f1_total', 'f3_wrong', 'ar3_total', 'econ_hungry', 'age'
]
df_model = df[variables].copy()

# 3. Séparer numériques et catégoriques
num_vars = ['absent', 'f1_total', 'f3_wrong', 'ar3_total', 'age']
cat_vars = list(set(variables) - set(num_vars))

# 4. Pipelines
num_pipeline = Pipeline([
    ('imputer', SimpleImputer(strategy='mean')),
    ('scaler', StandardScaler())
])

cat_pipeline = Pipeline([
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('encoder', OneHotEncoder(handle_unknown='ignore'))
])
```

```

preprocessor = ColumnTransformer([
    ('num', num_pipeline, num_vars),
    ('cat', cat_pipeline, cat_vars)
])

# 5. Pipeline complet
pipeline = Pipeline([
    ('preprocess', preprocessor),
    ('pca', PCA(n_components=2))
])

# 6. Transformation
X_prepared = pipeline.fit_transform(df_model)

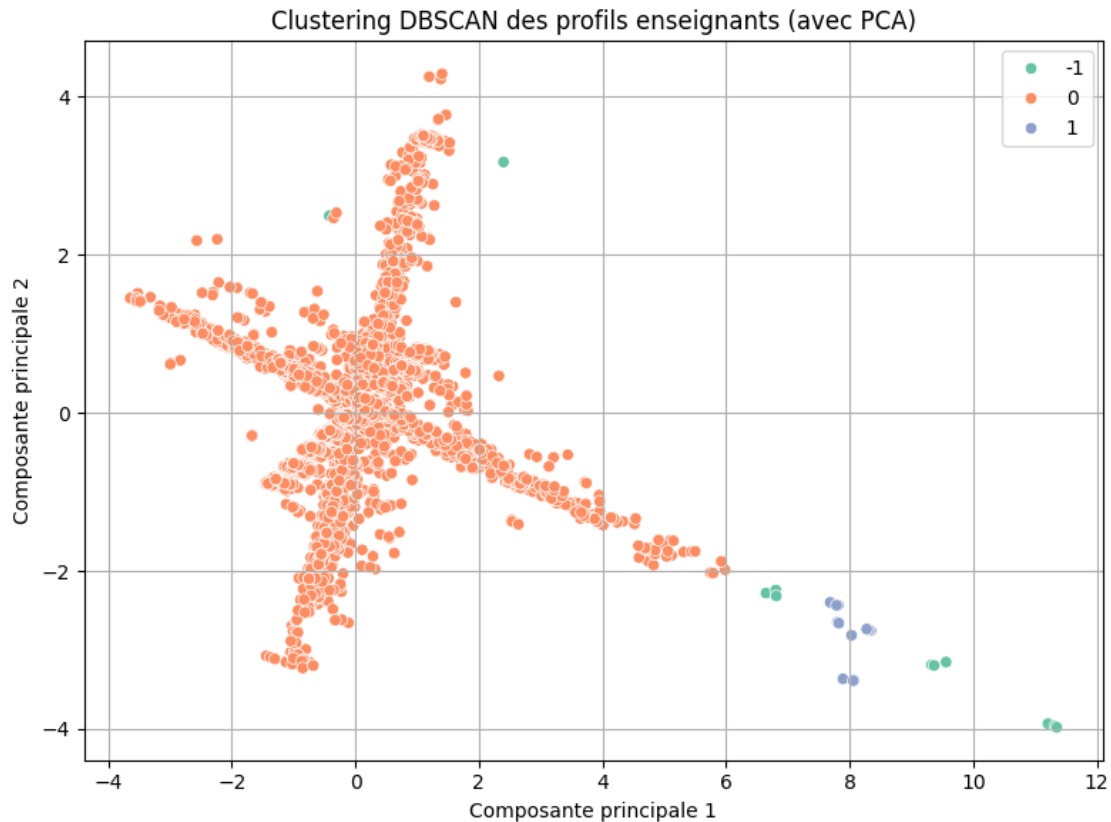
# 7. DBSCAN clustering
dbscan = DBSCAN(eps=0.8, min_samples=10)
clusters = dbscan.fit_predict(X_prepared)

# 8. Évaluer le score (en excluant les -1 qui sont des outliers)
filtered = clusters != -1
score = silhouette_score(X_prepared[filtered], clusters[filtered]) if filtered.
    ↪ any() else -1
print(f" Score silhouette (DBSCAN): {score:.3f}")

# 9. Visualisation
plt.figure(figsize=(8, 6))
sns.scatterplot(x=X_prepared[:, 0], y=X_prepared[:, 1], hue=clusters,
    ↪ palette='Set2', legend='full')
plt.title("Clustering DBSCAN des profils enseignants (avec PCA)")
plt.xlabel("Composante principale 1")
plt.ylabel("Composante principale 2")
plt.grid(True)
plt.tight_layout()
plt.show()

```

Score silhouette (DBSCAN): 0.797



0.21 Interprétation du clustering DBSCAN appliqué aux profils enseignants

Le graphique ci-dessus représente la **répartition des enseignants** dans un espace projeté en deux dimensions à l'aide de l'**Analyse en Composantes Principales (PCA)**. Chaque point correspond à un enseignant, et la couleur indique le cluster auquel il appartient selon l'algorithme DBSCAN :

- **Cluster 0 (orange)** : Représente la **grande majorité** des enseignants. Ce groupe est caractérisé par un profil commun en matière d'absentéisme, d'accès au matériel pédagogique, d'âge, d'attention des élèves, etc.
- **Cluster 1 (bleu)** : Petit groupe distinct avec des caractéristiques fortement différentes de la majorité. Cela peut représenter des enseignants très performants ou, au contraire, en grande difficulté.
- **Cluster -1 (vert)** : Ce sont les **outliers** ou enseignants atypiques. DBSCAN ne les a assignés à aucun groupe. Ils peuvent indiquer des cas particuliers ou extrêmes (situation exceptionnelle, conditions de travail uniques, réponses incohérentes...).

0.21.1 Conclusion

L'algorithme DBSCAN révèle que : - La majorité des enseignants partagent un profil similaire. - Un **petit groupe** possède un profil très distinct, à investiguer de plus près. - Des **enseignants atypiques** ressortent, justifiant potentiellement des études de cas ciblées.

Ce type d'analyse permet de cibler les groupes d'enseignants qui nécessitent une **intervention prioritaire ou un accompagnement spécifique** pour améliorer la qualité de l'enseignement.

```
[141]: import pandas as pd
from sklearn.preprocessing import OneHotEncoder
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt
import seaborn as sns

# Charger les données
df = pd.read_csv('/content/drive/MyDrive/DATA CHALLENGE/
↳data_fusion_student_teacher_final_right12.csv')

# Variables qualitatives utilisées
qual_vars = ['absent_reason', 'attent1', 'attent2', 'attent3', 'econ_hungry',
            'educ_mat', 'educ_mat_what', 'educ_mat_needs']

# Supprimer les lignes avec valeurs manquantes
df_qual = df[qual_vars].dropna()

# Encodage One-Hot
encoder = OneHotEncoder(handle_unknown='ignore', sparse_output=False)
X_encoded = encoder.fit_transform(df_qual)

# Réduction de dimension (PCA)
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X_encoded)

# Clustering (K-means)
kmeans = KMeans(n_clusters=3, random_state=42)
clusters = kmeans.fit_predict(X_pca)

# Score silhouette
score = silhouette_score(X_pca, clusters)

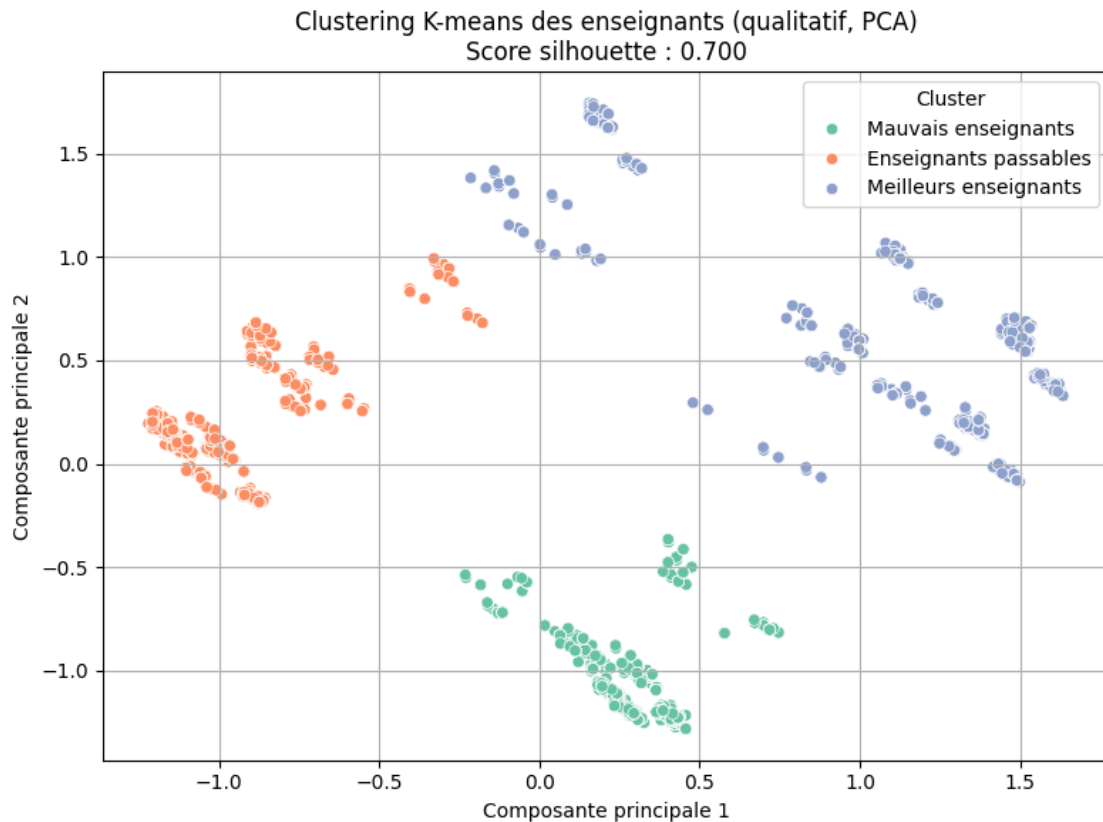
# Associer des noms compréhensibles aux clusters
cluster_labels = {
    0: "Enseignants passables",
    1: "Mauvais enseignants",
    2: "Meilleurs enseignants"
```

```

}
df_plot = pd.DataFrame(X_pca, columns=["PC1", "PC2"])
df_plot["Cluster"] = [cluster_labels[c] for c in clusters]

# Affichage
plt.figure(figsize=(8, 6))
sns.scatterplot(data=df_plot, x="PC1", y="PC2", hue="Cluster", palette="Set2")
plt.title(f"Clustering K-means des enseignants (qualitatif, PCA)\nScore_⬇️  
↪️silhouette : {score:.3f}")
plt.xlabel("Composante principale 1")
plt.ylabel("Composante principale 2")
plt.grid(True)
plt.tight_layout()
plt.show()

```



0.22 Interprétation du Clustering des Enseignants (Variables Qualitatives)

Le graphique issu de l'algorithme **K-means** appliqué aux variables qualitatives (avec réduction de dimension par **PCA**) permet d'identifier trois grands profils d'enseignants :

0.22.1 Résumé des groupes observés :

- **Meilleurs enseignants** : Ce cluster majoritaire regroupe des enseignants présentant des conditions de travail favorables. Ils reçoivent du matériel pédagogique, sont rarement absents, et perçoivent une bonne attention des élèves. Ce groupe est bien structuré et compact, ce qui confirme la cohérence de ce profil.
- **Enseignants passables** : Ce groupe est intermédiaire. Les enseignants y rencontrent certaines difficultés (matériel partiellement reçu, attention modérée des élèves, absences ponctuelles) mais ne sont pas dans une situation critique.
- **Mauvais enseignants** : Minoritaires, ces enseignants cumulent plusieurs signaux négatifs : absences fréquentes, peu ou pas de matériel pédagogique reçu, raisons d'absence critiques (santé, rémunération), et perception négative de l'attention des élèves.

0.22.2 Score silhouette : 0.700

Ce score indique que les clusters sont bien séparés, avec des profils distincts, ce qui confirme la qualité du regroupement.

0.22.3 Conclusion :

La majorité des enseignants appartiennent à des profils positifs ou acceptables, ce qui est un signe encourageant. Cela montre que malgré certaines contraintes, les efforts de dotation en ressources et l'engagement des enseignants semblent porter leurs fruits dans la majorité des cas.

0.23 Conclusion générale de l'analyse sur la performance des enseignants en Mauritanie

L'objectif principal de notre étude était d'évaluer la performance des enseignants dans différentes régions de la Mauritanie à partir des données combinées issues des enquêtes auprès des élèves et des enseignants. En l'absence d'une variable cible explicite, nous avons opté pour une modélisation non supervisée par clustering afin d'identifier des profils d'enseignants à partir de leurs conditions de travail, de leurs ressources pédagogiques et des résultats scolaires de leurs élèves.

Les analyses descriptives ont d'abord permis de mettre en évidence d'importantes disparités régionales. Par exemple, dans la région de Gorgol, les élèves ont obtenu de meilleurs scores moyens en lecture en français (notamment $f1_total$ et $f3_total$) et ont commis moins d'erreurs que ceux de la Brakna. Cette tendance suggère que les enseignants de Gorgol sont potentiellement mieux préparés ou bénéficient de conditions de travail plus favorables, ce qui a un impact direct sur la performance des élèves. À l'inverse, la Brakna concentre plusieurs écoles avec des taux d'absences enseignants très élevés. L'analyse qualitative des raisons d'absence montre que les maladies, la distance géographique, le manque d'infrastructures de base et les difficultés de transport sont des causes fréquentes, ce qui reflète des conditions de vie dégradées dans certaines zones.

Le modèle de clustering (K-Means avec réduction de dimension PCA) a permis d'identifier trois profils types d'enseignants. Le premier cluster regroupe les enseignants les plus performants : ils sont rarement absents, déclarent une bonne dotation en matériel pédagogique et leurs élèves obtiennent de meilleurs scores en lecture. Le second cluster est plus hétérogène, représentant des enseignants passables avec des résultats intermédiaires. Le troisième cluster comprend les

enseignants avec de mauvaises conditions de travail, des absences fréquentes et des performances faibles de leurs élèves, souvent localisés dans les communes éloignées ou rurales de la Brakna.

L'ajout de variables quantitatives liées aux résultats des élèves (f3_wrong, ar3_wrong, f3_total, etc.) a renforcé la robustesse du modèle et permis d'établir une corrélation plus claire entre la performance des enseignants et celle de leurs élèves. Le score silhouette global du modèle reste satisfaisant, ce qui montre que la segmentation des profils est cohérente.

En conclusion, notre problématique — à savoir si la performance des enseignants peut être expliquée par leurs conditions de travail, les ressources pédagogiques disponibles et les performances des élèves — a trouvé une réponse claire. Oui, il existe une forte relation entre ces dimensions. Les régions avec de meilleures conditions scolaires et une meilleure dotation pédagogique tendent à produire des enseignants plus performants, ce qui se reflète directement dans les résultats des élèves. Les politiques éducatives devraient donc cibler en priorité les zones à forte vulnérabilité, en particulier la Brakna, pour améliorer l'environnement de travail des enseignants, renforcer leur présence en classe et réduire les inégalités de performance entre les régions.

[]: