IUT de Paris - Rives de Seine
Université Paris Cité

Pour le 08/01/2024

À Amine Meddour

Projet : Performances des Étudiants

Aya El Yaouti
BUT 2 SD PARCOURS VCOD

Sommaire :

I) Base de données.

a) Description de la base de données

Dans ce rapport, nous allons nous intéresser à une base de données qui étudie les performances des étudiants dans différentes matières telles que les mathématiques, la lecture et l'écriture. Nous allons également prendre en compte des indicateurs tels que la préparation aux examens, le genre et le niveau d'éducation des parents.
Notre base de données est un fichier csv (PerformancesStudents.csv), qui contient 8 variables distinctes qui sont :

Math Score : Note en mathématiques
Reading Score : Note en lecture
Writing Score : Note en écriture
Gender : Genre (homme et femme)
Race/Ethnicity : Race ou origine ethnique (Groupe A,B,C,D,E)
Parental level of education : Niveau d'éducation des parents
Lunch : Déjeuner. Standard est pour les personnes payant le déjeuner à l'école, tandis que free/reduced (gratuit/réduit) est pour les personnes dont le revenu familial est inférieur de 130% à 185% du seuil de pauvreté aux États-Unis. Cela indique le niveau économique de l'étudiant.
Test preparation course : Cours de préparation aux examens. Peut être complété (completed) ou non réalisé (none).

Avec toutes ces informations, nous cherchons à mieux comprendre les facteurs qui influent sur les résultats des étudiants. En examinant ces données, nous pourrons identifier des modèles et des tendances significatifs, et espérons proposer des idées pour aider les étudiants à s'améliorer à l'école.

PROBLÉMATIQUE :

Comment mettre en place des actions ciblées pour réduire les écarts de genre, soutenir les étudiants issus de milieux éducatifs moins favorisés, favoriser l'équité entre les groupes ethniques, améliorer l'accès à une alimentation équilibrée, et encourager la participation aux cours de préparation aux tests en vue d'optimiser les performances académiques des étudiants ?

b) Exploration de la base de données

Le nombre de valeurs nulles dans chaque colonne du DataFrame (df) ainsi que le type de données associé à chaque variable :

```
gender                         0     gender                         object
race/ethnicity                 0     race/ethnicity                 object
parental level of education    0     parental level of education    object
lunch                          0     lunch                          object
test preparation course        0     test preparation course        object
math score                     0     math score                      int64
reading score                  0     reading score                   int64
writing score                  0     writing score                   int64
dtype: int64                         dtype: object
```

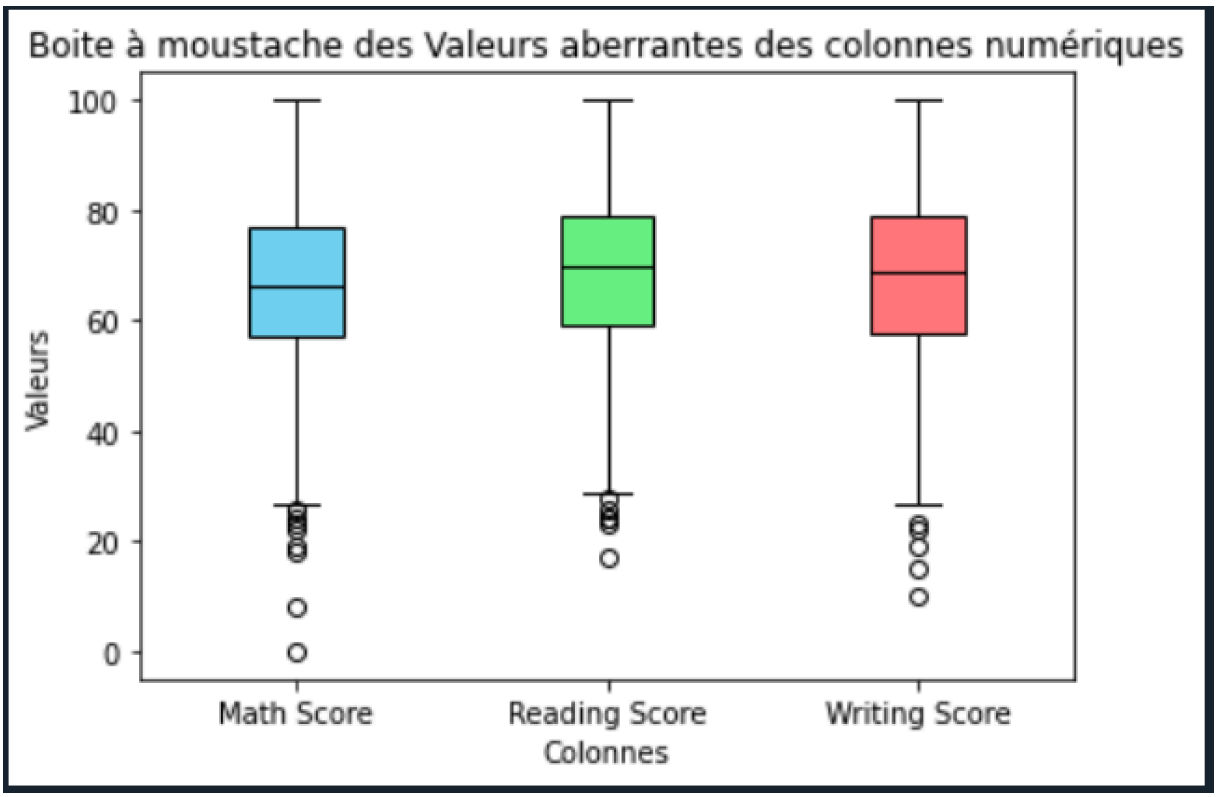
Distribution des valeurs dans certaines colonnes de notre base de données sous forme de pourcentages :

```
dtype: int64
gender
female    51.8
male      48.2
Name: proportion, dtype: float64
race/ethnicity
group C    31.9
group D    26.2
group B    19.0
group E    14.0
group A     8.9
Name: proportion, dtype: float64
parental level of education
some college          22.6
associate's degree    22.2
high school           19.6
some high school      17.9
bachelor's degree     11.8
master's degree        5.9
Name: proportion, dtype: float64
lunch
standard        64.5
free/reduced    35.5
Name: proportion, dtype: float64
test preparation course
none         64.2
completed    35.8
Name: proportion, dtype: float64
```

L'ensemble de données semble complet, sans aucune valeur manquante dans toutes les colonnes. Les variables comprennent des caractéristiques catégorielles telles que le genre, l'origine ethnique, le niveau d'éducation des parents, le type de déjeuner, et la participation à un cours de préparation au test. Des variables numériques comprennent les scores en mathématiques, lecture et écriture.

Concernant la distribution des données, la répartition entre les genres est équilibrée, avec une proportion de 51,8% de femmes et 48,2% d'hommes. En ce qui concerne l'origine ethnique, le groupe C est le plus représenté avec 31,9%, suivi du groupe D avec 26,2% et du groupe B avec 19,0%. Pour le niveau d'éducation des parents, environ 22,6% ont suivi des cours universitaires, 22,2% ont obtenu un diplôme associé, et 11,8% ont un diplôme de bachelier. Concernant le déjeuner, la majorité (64,5%) des étudiants ont un déjeuner "standard", tandis que 35,5% bénéficient d'un déjeuner "gratuit/réduit". Enfin, en ce qui concerne le cours de préparation au test, la majorité (64,2%) des étudiants n'ont suivi aucun cours, tandis que 35,8% ont complété le cours de préparation.

	math score	reading score	writing score
count	1000.00000	1000.000000	1000.000000
mean	66.08900	69.169000	68.054000
std	15.16308	14.600192	15.195657
min	0.00000	17.000000	10.000000
25%	57.00000	59.000000	57.750000
50%	66.00000	70.000000	69.000000
75%	77.00000	79.000000	79.000000
max	100.00000	100.000000	100.000000



En ce qui concerne les scores des épreuves, on observe quelques statistiques descriptives intéressantes. La moyenne des scores en mathématiques est d'environ 66,09, pour la lecture c'est d'environ 69,17, et pour l'écriture, elle est d'environ 68,05. Ces moyennes donnent une idée générale des performances des étudiants dans chaque épreuve.

En ce qui concerne les valeurs aberrantes, on peut remarquer quelques points à considérer. Tout d'abord, le score minimum de 0 en mathématiques pourrait être considéré comme une valeur aberrante, car il est assez éloigné de la moyenne et des quartiles. De même, le score minimum de 17 en lecture et de 10 en écriture peut également être considéré comme atypique.

L'écart type fournit une mesure de la dispersion des scores autour de la moyenne. Des écarts types relativement élevés (15,16 en mathématiques, 14,60 en lecture et 15,20 en écriture) indiquent une certaine variabilité des performances entre les étudiants.

En examinant le tableau des valeurs minimales (min) et le premier quartile (25%), on peut identifier les valeurs potentiellement aberrantes. Les scores en mathématiques semblent avoir une distribution plus étendue, avec une valeur minimale de 0 et un premier quartile à

57, tandis que les scores en lecture et écriture semblent plus concentrés autour de valeurs plus élevées.

c) Impact de la Suppression des Valeurs Aberrantes sur les Performances des Étudiants

```
La forme du DataFrame avant la suppression des valeurs aberrantes est(1000, 8)
La forme du DataFrame après la suppression des valeurs aberrantes est (988, 8)
```

	math score	reading score	writing score
count	988.000000	988.000000	988.000000
mean	66.625506	69.640688	68.566802
std	14.409394	14.016760	14.525267
min	27.000000	29.000000	27.000000
25%	57.000000	60.000000	58.000000
50%	66.000000	70.000000	69.000000
75%	77.000000	80.000000	79.000000
max	100.000000	100.000000	100.000000

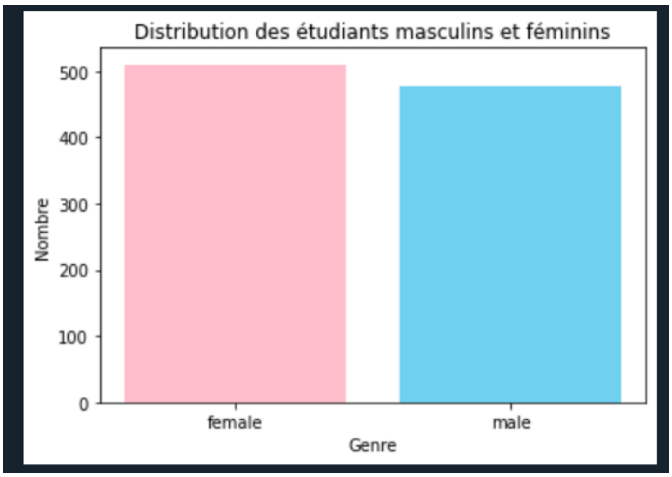
Le DataFrame initial avait une forme de (1000, 8) avant la suppression des valeurs aberrantes. Après cette opération, la forme du DataFrame est passée à (988, 8). La suppression des valeurs aberrantes a permis d'éliminer certaines observations atypiques, influençant ainsi les statistiques descriptives des scores en mathématiques, lecture et écriture.

La moyenne des scores en mathématiques, lecture et écriture est maintenant de 66,63, 69,64 et 68,57 respectivement. Les écarts types ont également été ajustés, passant à 14,41 pour les mathématiques, 14,02 pour la lecture, et 14,53 pour l'écriture.

Les valeurs aberrantes ont été éliminées, modifiant les statistiques descriptives pour mieux refléter la tendance centrale et la dispersion des performances des étudiants. Ce processus contribue à une compréhension plus précise et fiable des résultats scolaires.

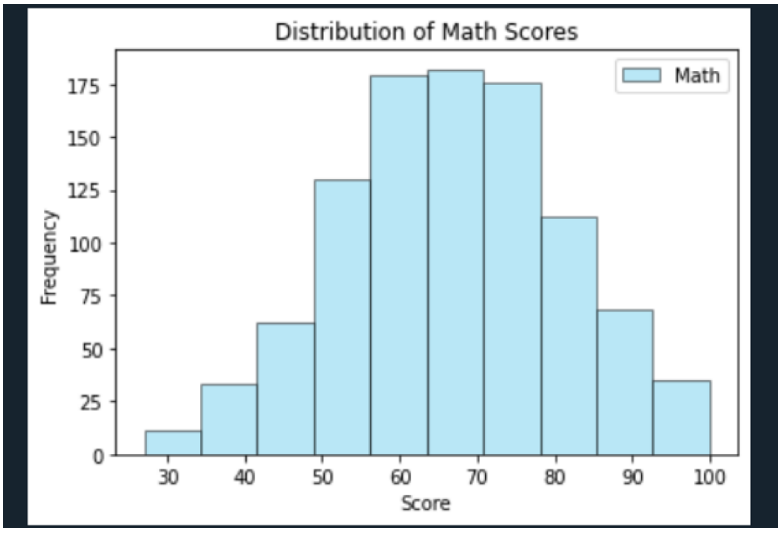
II) Réalisation de la visualisation des données

a) Distribution des étudiants par genre

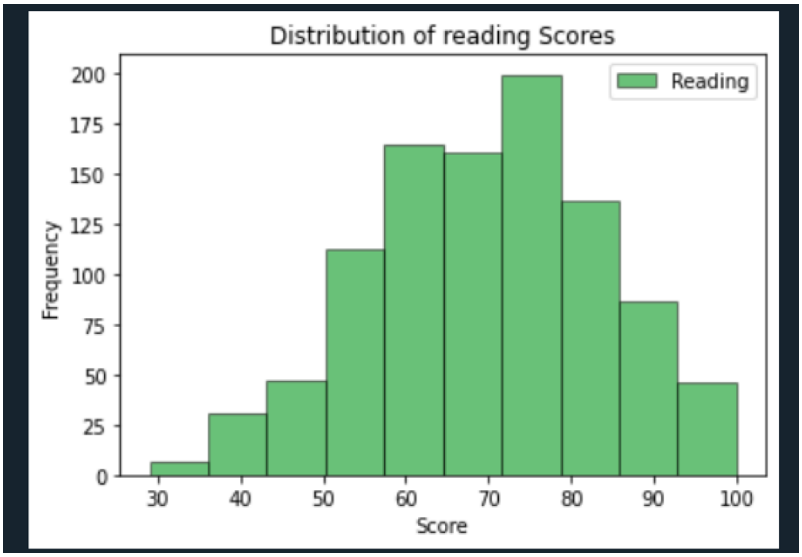


Nous pouvons voir qu'il y a plus de femmes que d'hommes.

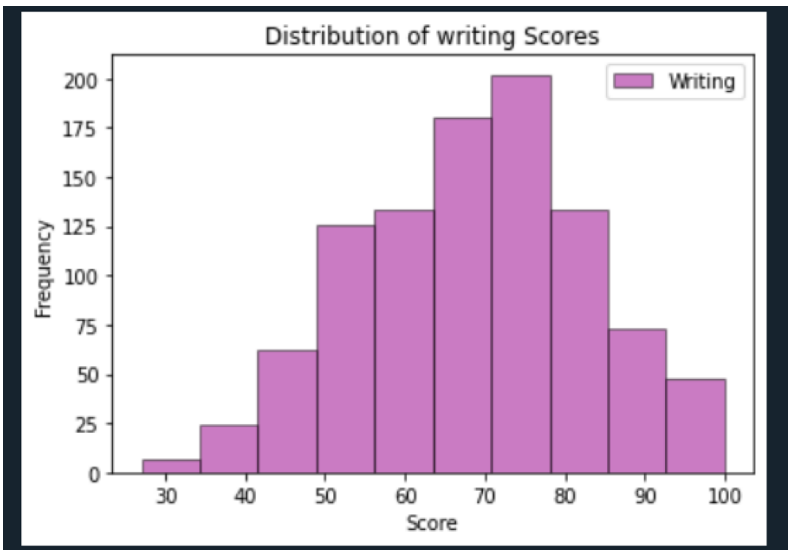
b) Distribution des scores en Mathématiques, Lecture et Écriture



La répartition des scores en mathématiques est globalement symétrique, avec une concentration marquée entre 60 et 70. Le mode, qui représente la valeur la plus fréquemment observée, est spécifiquement situé à 65. Cela suggère que la plupart des étudiants ont obtenu des scores proches de cette valeur, indiquant une tendance centrale autour de ce point.

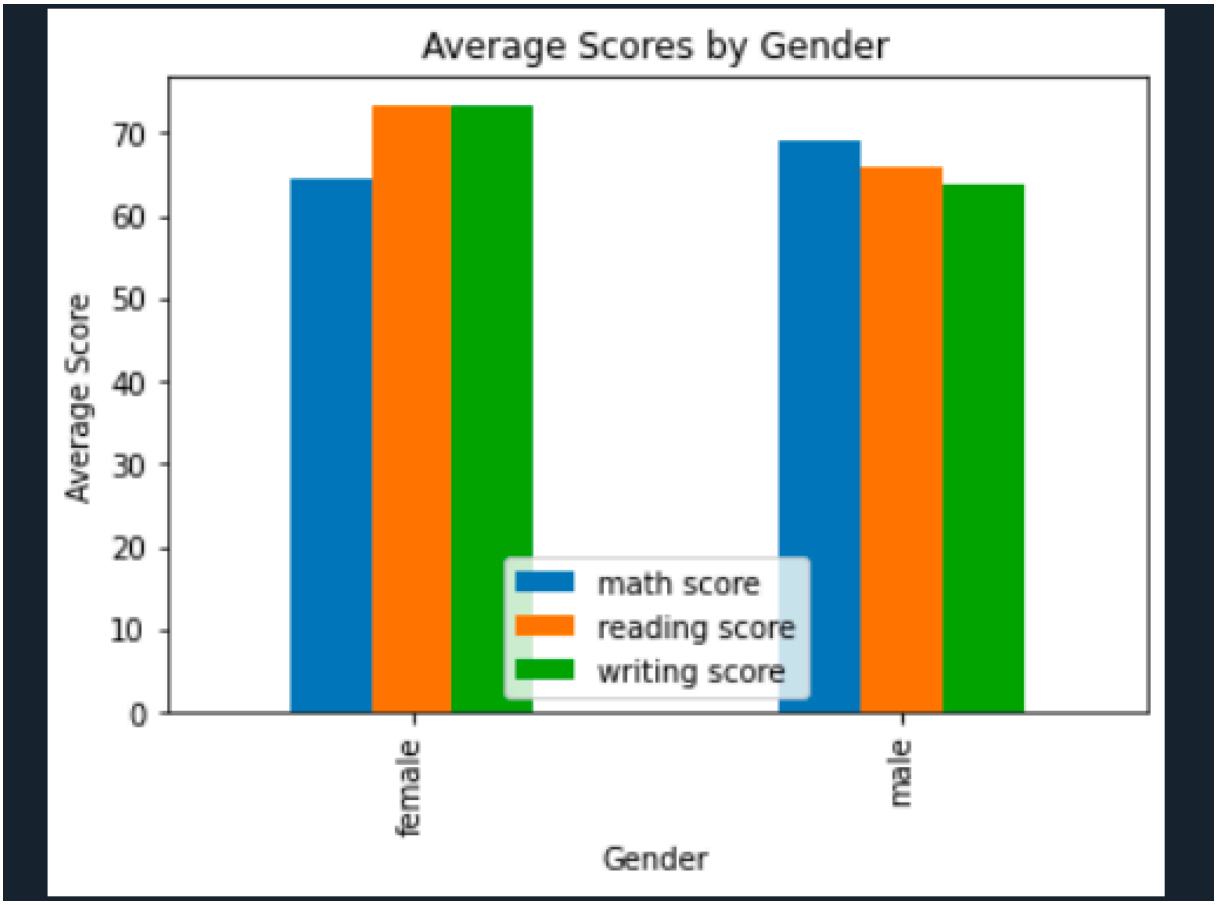


La répartition des scores en lecture est globalement symétrique, avec une majorité de scores compris entre 60 et 75. Le mode, représenté par la valeur la plus fréquemment observée, est spécifiquement établi à 75. Cette concentration autour de 75 indique une tendance centrale marquée, suggérant que de nombreux étudiants ont obtenu des scores proches de cette valeur, renforçant ainsi la symétrie de la distribution.



La distribution des scores en lecture, bien qu'approximativement symétrique, présente une concentration notable entre 60 et 80, avec une valeur modale de 70, suggérant que de nombreux étudiants ont obtenu des scores proches de cette valeur, renforçant ainsi la symétrie de la distribution.

c) Scores moyens par genre

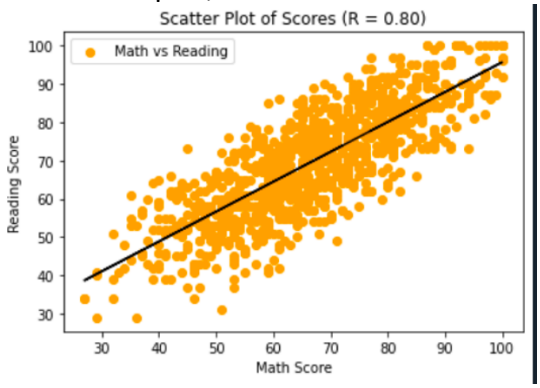


En général, nous pouvons observer que les femmes obtiennent des scores plus élevés en lecture et écriture, tandis que les hommes obtiennent des scores plus élevés en mathématiques.

III) Corrélation

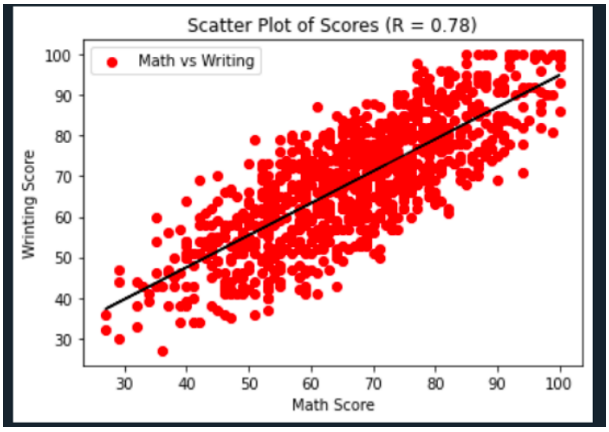
a) Corrélation entre les scores en maths et en lecture

Création de graphiques de dispersion pour montrer la relation entre les scores en mathématiques, en lecture et en écriture.



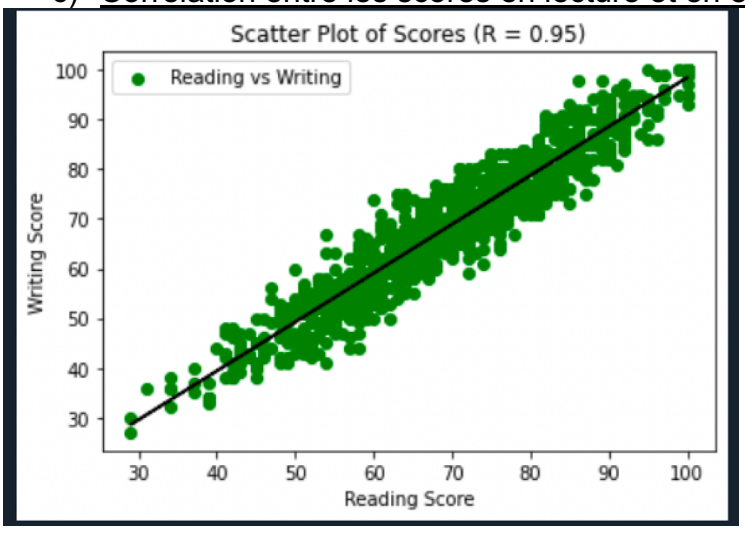
Le score en mathématiques et les scores en lecture présentent une corrélation positive entre eux, avec un coefficient de corrélation (R) de 0,80. En d'autres termes, les étudiants obtenant de bons résultats en mathématiques ont également tendance à excelller en lecture, et vice versa.

b) Corrélation entre les scores maths et en écriture



Le score en mathématiques et les scores en écriture présentent une corrélation positive entre eux, avec un coefficient de corrélation (R) de 0,78. Cette corrélation renforce l'idée que les étudiants réussissant bien en mathématiques ont également tendance à excelller en écriture, et vice versa.

c) Corrélation entre les scores en lecture et en écriture



Les scores en écriture et les scores en lecture ont une corrélation positive entre eux, avec une valeur de R égale à 0,95. Cette valeur élevée de R suggère que les étudiants obtenant de bons résultats en écriture ont également tendance à excelller en lecture, et vice versa, soulignant une forte corrélation entre ces deux compétences

IV) Analyse statistique de base.

a) Statistiques descriptives des scores en math, lecture et écriture

```
Mean Scores:
math score       66.63
reading score    69.64
writing score    68.57
dtype: float64

Median Scores:
math score       66.0
reading score    70.0
writing score    69.0
dtype: float64

Mode Scores:
   math score  reading score  writing score
0          65             72             74
```

Les scores moyens en mathématiques, lecture et écriture sont respectivement de 66.63, 69.64 et 68.57. Les médianes correspondantes sont de 66.0, 70.0 et 69.0. Les modes, représentés par les valeurs les plus fréquemment observées, sont de 65 en mathématiques, 72 en lecture et 74 en écriture.

b) Écart-type des Scores en Math, Lecture et Écriture.

```
                Standard Deviation
math score                   14.41
reading score                14.02
writing score                14.53
```

L'écart-type mesure à quel point les scores se dispersent autour de la moyenne. Pour les mathématiques, il est de 14.41, pour la lecture de 14.02, et pour l'écriture de 14.53. Ces valeurs élevées suggèrent que les résultats des étudiants varient considérablement. Les scores sont répartis sur une large plage par rapport à la moyenne, indiquant une diversité notable dans les performances académiques.

V) Test : Évaluation des performances entre sexes en math, lecture et écriture

a) En mathématiques

Hypothèse nulle : Il n'y a pas de différence significative entre les performances en mathématiques des étudiants de sexe masculin et féminin.

Hypothèse alternative : Il existe une différence significative entre les performances en mathématiques des étudiants de sexe masculin et féminin.

```
we reject the null hypothesis
p-value =  2.640262881425632e-07
t statistic =  5.183610882116533
```

Nous rejetons l'hypothèse nulle, il existe bien une différence significative entre les performances en mathématiques des étudiants de sexe masculin et féminin.

b) En lecture

Hypothèse nulle : Il n'y a pas de différence significative entre les performances en lecture des étudiants de sexe masculin et féminin.

Hypothèse alternative : Il existe une différence significative entre les performances en lecture des étudiants de sexe masculin et féminin.

```
we reject the null hypothesis
p-value =  2.6052257840994235e-17
t statistic =  -8.621903833772542
```

Nous rejetons l'hypothèse nulle, il existe bien une différence significative entre les performances en lecture des étudiants de sexe masculin et féminin.

c) En écriture

Hypothèse nulle : Il n'y a pas de différence significative entre les performances en écriture des étudiants de sexe masculin et féminin.

Hypothèse alternative : Il existe une différence significative entre les performances en écriture des étudiants de sexe masculin et féminin.

```
we reject the null hypothesis
p-value =  4.970687182539877e-26
t statistic =  -8.621903833772542
```

Nous rejetons l'hypothèse nulle, il existe bien une différence significative entre les performances en écriture des étudiants de sexe masculin et féminin.

VI) Effet d'un Cours de Préparation aux Tests sur les résultats des Étudiants

```
              Mean Scores (Completed course)
math score                             69.83
reading score                          73.98
writing score                          74.53
              Mean Scores (did not complete course)
math score                                64.81
reading score                             67.19
writing score                             65.19
```

Les étudiants qui ont suivi le cours de préparation à l'examen ont obtenu de meilleurs résultats globaux (en mathématiques, en lecture et en écriture).

Réalisation d'un test T :

Hypothèse nulle : Les scores en mathématiques des étudiants ayant suivi le cours de préparation aux tests ne sont pas différents de la distribution de la population.

Hypothèse alternative : Les échantillons de scores en mathématiques des étudiants ayant suivi le test sont différents de la population. Réalisons un test t au niveau de confiance de 95% et de 99% pour voir s'il rejette correctement l'hypothèse nulle selon laquelle l'échantillon provient de la même distribution que la population.

a) Test de Student 95%

Statistique t et p-valeur :

```
4.243733605109862 2.8073583196986977e-05
```

Mise en place des quantiles pour comparaison avec la statistique t :

```
-1.9666499952118222 1.9666499952118217
```

Calcul de l'intervalle de confiance

```
66.6255060728745
(67.84453005396585, 71.8081310104599)
```

La statistique t est de 4.24 avec une p-valeur de 0.0000281, indiquant un rejet significatif de l'hypothèse nulle.

L'intervalle de confiance à 95% pour la moyenne des scores en mathématiques (67.84 à 71.81) exclut la moyenne nulle. Ainsi, nous avons des preuves solides que les scores en mathématiques des étudiants ayant suivi le cours de préparation aux tests diffèrent de la population générale

b) Test de Student 99%

```
66.6255060728745
(67.87304899406075, 71.779612070365)
```

Au niveau de confiance de 99%, nous rejetons l'hypothèse nulle.

La moyenne des scores en mathématiques est de 66.63, et l'intervalle de confiance à 99% pour cette moyenne est de 67.87 à 71.78. Puisque la moyenne observée de 66.63 est incluse dans cet intervalle, nous rejetons l'hypothèse nulle au niveau de confiance de 99%. Cela suggère de manière significative que les scores en mathématiques des étudiants ayant suivi le cours de préparation aux tests diffèrent de la population générale.

VII) Analyse des Facteurs Influents sur les Scores Académiques : Une Étude Comparative

a) Découvrir s'il existe une différence entre les scores des hommes et des femmes.

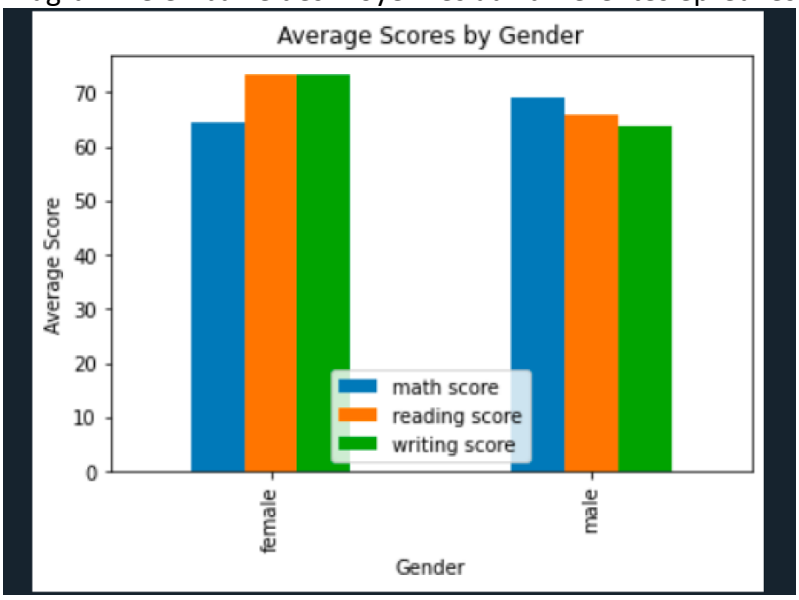
Moyenne aux épreuves par sexe :

	female	male
math score	64.36	69.05
reading score	73.23	65.81
writing score	73.16	63.67

Moyenne aux épreuves des hommes et femmes ayant ou n'ayant pas eu de cours de préparation :

```
                Mean Scores (Male Completed)
math score                           72.34
reading score                        70.21
writing score                        69.79
                Mean Scores (Female Completed)
math score                           67.44
reading score                        77.56
writing score                        79.03
                Mean Scores (Male not completed)
math score                           67.16
reading score                        63.29
writing score                        60.16
                Mean Scores (Female not completed)
math score                           62.63
reading score                        70.81
writing score                        69.87
```

Diagramme en barre des moyennes aux différentes épreuves par sexe :



Les hommes ont surpassé les femmes en mathématiques, tandis que les femmes ont surpassé les hommes à la fois en lecture et en écriture.

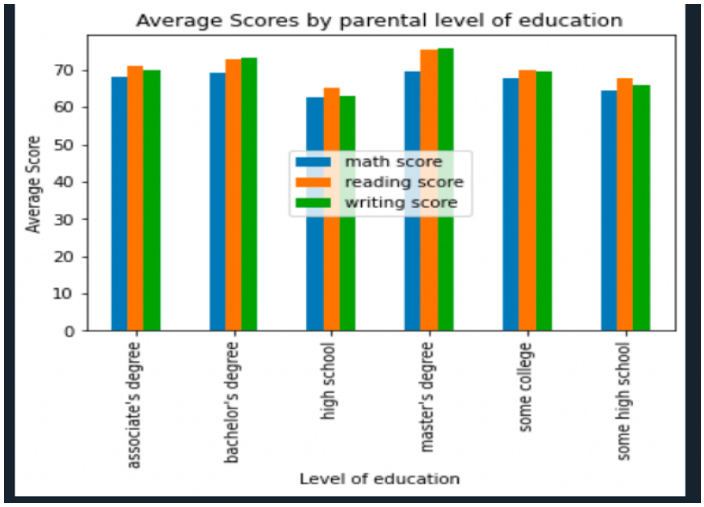
Les hommes qui ont suivi le cours de préparation à l'examen ont obtenu de meilleurs résultats en mathématiques que les filles qui ont suivi le cours, mais ont obtenu de moins bons résultats en lecture et en écriture.

Les hommes qui n'ont pas suivi le cours de préparation à l'examen ont obtenu de meilleurs résultats en mathématiques que les femmes qui n'ont pas suivi le cours, mais ont obtenu de moins bons résultats en lecture et en écriture.

Tant les hommes que les femmes qui ont suivi le cours de préparation ont obtenu de meilleurs résultats aux trois tests que ceux qui ne l'ont pas suivi.

Les hommes qui n'ont pas suivi le cours et les femmes qui l'ont suivi ont obtenu les mêmes résultats en mathématiques.

b) Est-ce que le niveau d'éducation des parents influence les scores ?



ANOVA test :

Hypothèse nulle : Les distributions des scores en écriture ne diffèrent pas entre les groupes selon le niveau d'éducation des parents.

```
5.685043306239461e-13
we reject the null hypothesis
```

Les résultats des tests des étudiants ont tendance à baisser lorsque leurs parents ont un niveau d'éducation inférieur à d'autres. Cependant, les étudiants étiquetés comme ayant "quelques années de lycée" ont obtenu des scores plus élevés que ceux étiquetés comme ayant "terminé le lycée".

c) Est-ce que les groupes ethniques ont un impact sur les scores ?

Moyenne aux différentes épreuves par groupe ethnique

	group B	group C	group A	group D	group E
math score	64.86	64.90	62.01	67.52	74.14
reading score	68.46	69.49	65.15	70.18	73.37
writing score	66.86	68.25	63.17	70.27	71.76

Les corrélations calculées pour chaque colonne de score.

```
{'math score': 0.21, 'reading score': 0.14, 'writing score': 0.16}
```

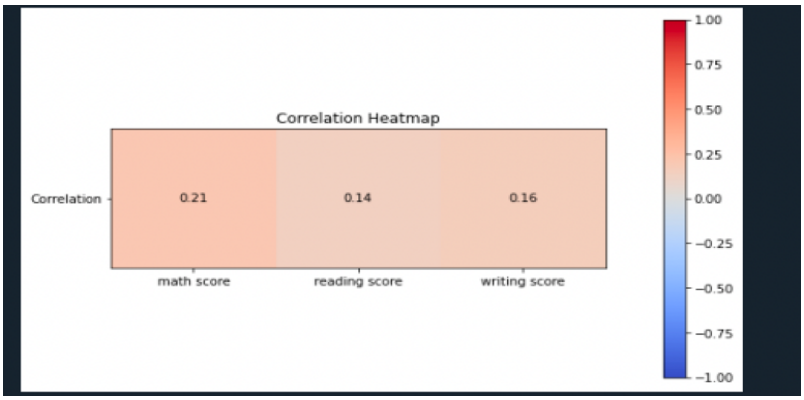
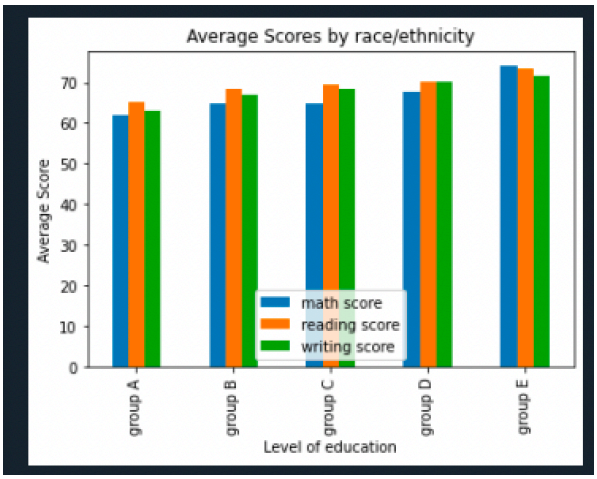


Diagramme en barre du scores aux différentes épreuves par groupe ethnique



Les scores moyens des groupes ethniques dans les trois matières sont les suivants : le Groupe E est le plus élevé, suivi par le Groupe D, le Groupe C, le Groupe B et le Groupe A.

d) Est-ce que le type de déjeuner a un impact sur les scores ?

Moyenne aux épreuves den fonction du type de déjeuner

	standard	free/reduced
math score	70.18	60.01
reading score	71.78	65.66
writing score	70.96	64.11

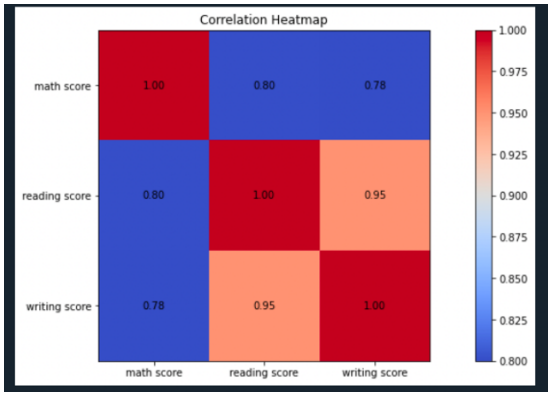
```
                                Standard and Completed
math score                                      73.53
reading score                                   76.22
writing score                                   76.77
                                Free and Completed
math score                                63.35
reading score                             70.07
writing score                             70.62
                                Standard and not completed
math score                                      68.34
reading score                                   69.36
writing score                                   67.79
                                Free and not completed
math score                                      57.99
reading score                                   62.99
writing score                                   60.17
```

Les étudiants qui ont pris un déjeuner régulier ont surpassé ceux qui ont pris un déjeuner gratuit ou à tarif réduit.

Les étudiants qui ont pris un déjeuner standard et ont suivi le cours de préparation à l'examen ont obtenu la moyenne la plus élevée par rapport aux étudiants qui ont pris un déjeuner standard et n'ont pas suivi le cours, aux étudiants qui ont pris un déjeuner gratuit et n'ont pas suivi le cours, et aux étudiants qui ont pris un déjeuner gratuit et n'ont pas suivi le cours. Cependant, les élèves qui ont pris un déjeuner gratuit et ont suivi le cours ont obtenu de meilleurs résultats en lecture et en écriture que ceux qui ont pris un déjeuner normal mais n'ont pas suivi le cours.

VII) Corrélation

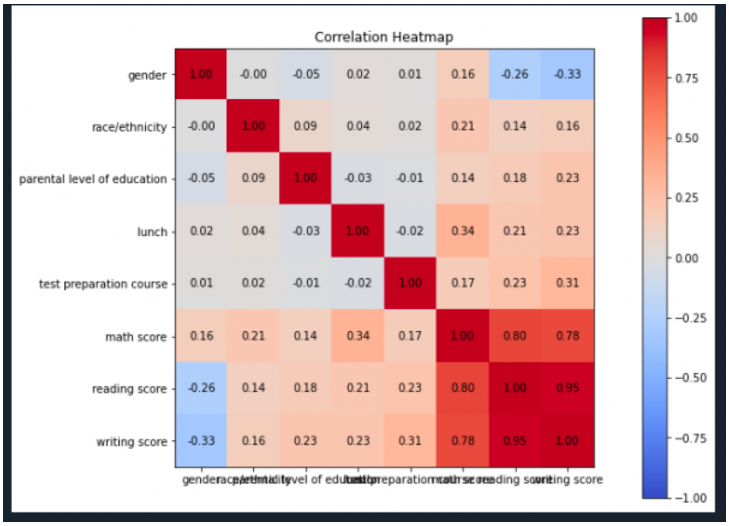
a) La corrélation des scores.



Les scores présentent une forte corrélation positive entre eux. Cela signifie que les élèves qui excellent dans une matière ont très probablement de bons résultats dans une autre matière.

b) Caractéristiques Clés pour l'Entraînement d'un Modèle ML

Quelles caractéristiques sont importantes pour entraîner un modèle d'apprentissage automatique ?



Selon le tableau de la matrice de corrélation, nous pouvons conclure que chaque colonne a un effet sur les trois scores de test. Par conséquent, nous ne pouvons pas ignorer ou supprimer l'une des colonnes lors de l'entraînement d'un modèle d'apprentissage automatique pour prédire les scores.

IX) Conclusion.

En conclusion, cette analyse approfondie des performances des étudiants met en lumière plusieurs facteurs influents sur les scores académiques. Certains points clés à retenir sont les suivants :

1. Impact du Genre : Des différences significatives entre les performances des hommes et des femmes ont été observées, avec des tendances spécifiques dans chaque matière. Il est crucial de comprendre et d'adresser ces disparités pour promouvoir l'équité éducative.

2. Influence du Niveau d'Éducation des Parents : Une corrélation entre le niveau d'éducation des parents et les performances académiques a été établie. Des efforts visant à soutenir les étudiants dont les parents ont un niveau d'éducation plus bas peuvent contribuer à améliorer les résultats scolaires.

3. Effet des Groupes Ethniques : Des variations dans les scores entre les groupes ethniques ont été identifiées. Une approche inclusive et des programmes spécifiques peuvent être développés pour soutenir les groupes qui montrent des performances plus faibles.

4. Influence du Type de Déjeuner : Le type de déjeuner a été lié aux performances académiques. Des initiatives pour améliorer l'accès à des repas nutritifs peuvent avoir un impact positif sur les résultats scolaires.

5. Répercussions du Cours de Préparation aux Tests : Les résultats montrent que les étudiants ayant suivi un cours de préparation aux tests ont obtenu des scores plus élevés. Encourager la participation à de tels cours peut être bénéfique.

a) Propositions pour l'Amélioration

1. Programmes de Soutien : Mettre en place des programmes de soutien éducatif pour les étudiants en fonction de leur genre, du niveau d'éducation des parents et de l'origine ethnique.

2. Équité d'Accès : Garantir un accès équitable à des ressources éducatives, y compris des repas nutritifs, pour tous les étudiants, indépendamment de leur situation économique.

3. Sensibilisation et Formation : Sensibiliser les enseignants, les parents et les élèves sur les différences de genre et les implications sur l'éducation, tout en fournissant des formations pour favoriser un environnement d'apprentissage inclusif.

4. Renforcement des Cours de Préparation aux Tests : Encourager la participation aux cours de préparation aux tests et développer des programmes plus accessibles pour tous les étudiants.

5. Suivi Continu : Établir un suivi continu des performances académiques, en identifiant rapidement les étudiants en difficulté et en mettant en place des interventions ciblées.

En adoptant ces propositions, les établissements scolaires peuvent œuvrer vers une éducation plus inclusive et équitable, favorisant le succès académique de tous les étudiants.