projet 8

# Communiquez les résultats

## Analyser les facteurs influents la performance des élèves aux examens

ACHAT HAYAT

# Sommaire

# Introduction

L'éducation est la clef du développement. La question de la réussite des élèves est au cœur de l'actualité et des préoccupations de notre société. Il est intéressant d'analyser les facteurs influençant leurs réussites, et ce en étudiant leurs natures, l'environnement auquel ils appartiennent ainsi que les conditions auxquelles ils sont confrontés.

C'est pour cela que j'ai opté pour le thème "analyser la performance des élèves aux examens", parmi tant d'autres que j'ai trouvé sur le site Kagel ou j'ai pu récupérer une base de données publique, comprenant des données de 1000 élèves d'un lycée situé au États Unies : sur leurs genre, appartenance ethnique , le niveau des parents, leurs préparation aux examens, ainsi que leurs scores(en math, en écrit et en lecture) .

Pour mener à bien mes analyses, j' ai commencé par explorer les données en vérifiant leurs justesses, puis en analysant chaque variable et leurs relations ainsi que le taux d'influence de chaque variable sur les résultats. Enfin, partitionner les élèves tout en les classant et analysant les clusters .

# 1 Exploration des données

j'ai chargé la base de données sur mon notebook à l'aide de la commande read.csv pour obtenir un dataframe comportant 1000 entrées et 8 colonnes:

| | gender | race/ethnicity | parental level of education | lunch | test preparation course | math score | reading score | writing score |
|---|---|---|---|---|---|---|---|---|
| 0 | female | group B | bachelor's degree | standard | none | 72 | 72 | 74 |
| 1 | female | group C | some college | standard | completed | 69 | 90 | 88 |
| 2 | female | group B | master's degree | standard | none | 90 | 95 | 93 |
| 3 | male | group A | associate's degree | free/reduced | none | 47 | 57 | 44 |

- Gender:le genre "mâle", "femelle".
- Race/ethnicity: l'appartenance ethnique 5 groupes: A,B,C,D,E.
- Parental level of education:niveau des parents:'niveau master' , 'niveau bachelor, 'associé','niveau bac ', 'études secondaires ' , 'niveau collège'.
- Lunch:'standard', 'free/reduced'.
- Test préparation course:cours de préparation aux tests:'none', 'completed'.
- Math score: score en mathématique.
- Reading score: score en lecture.
- Writing score: score en écriture.

## 1.1 Nettoyage des données

les données sont complètes, ne comportant ni de données manquantes ni de nulles. Pour toutes les variables, les valeurs semblent logiques et n'ont pas de valeurs aberrantes ce qui ne nécessite pas de nettoyage, donc on peut partir sur ce data frame .

```
: stu_perf.describe()
```

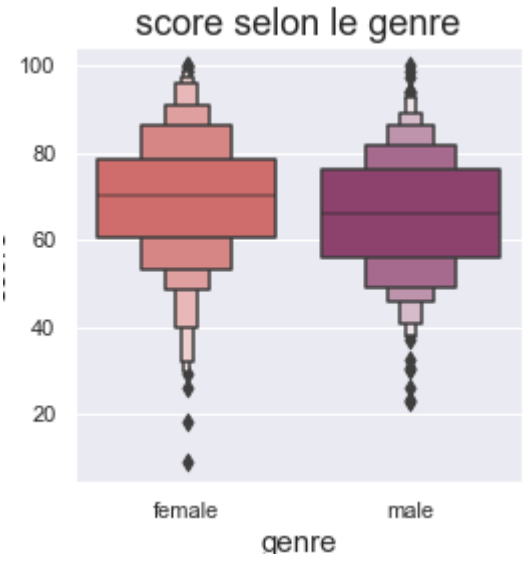
| | math score | reading score | writing score |
|---|---|---|---|
| count | 1000.00000 | 1000.000000 | 1000.000000 |
| mean | 66.08900 | 69.169000 | 68.054000 |
| std | 15.16308 | 14.600192 | 15.195657 |
| min | 0.00000 | 17.000000 | 10.000000 |
| 25% | 57.00000 | 59.000000 | 57.750000 |
| 50% | 66.00000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

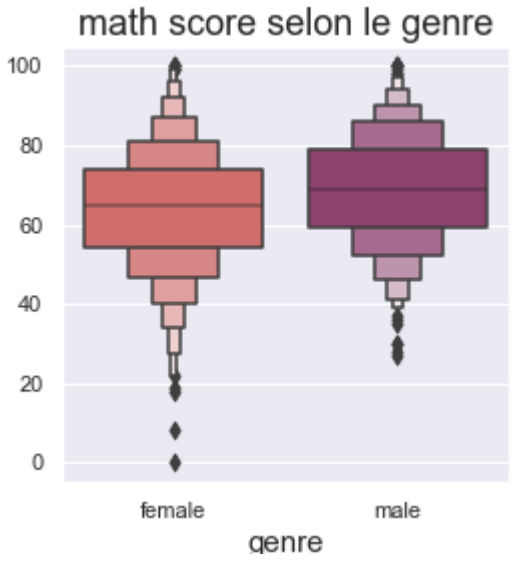
# 2 Analyses multivariées

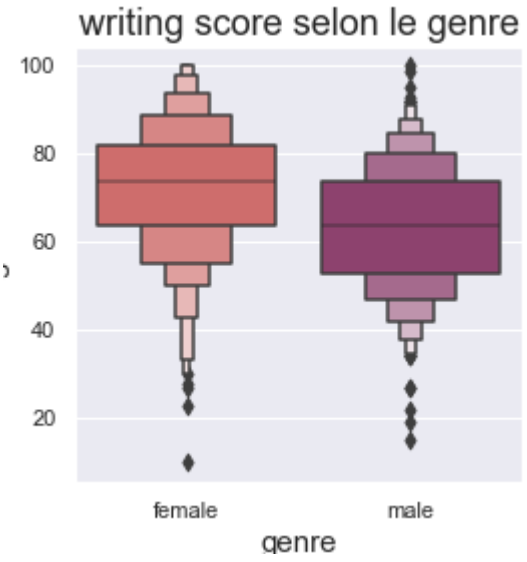
## 2.1 Analyse des résultats selon le genre

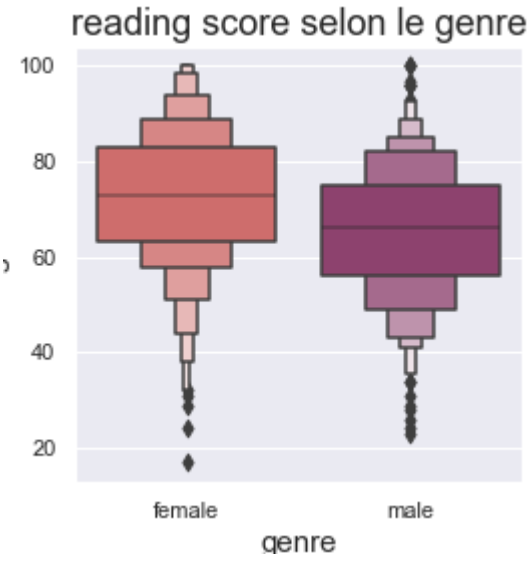
la part des femmes est un peu plus importante que les hommes,

une disparité de niveau entre les deux sexes , les femmes qui est plus élevée .



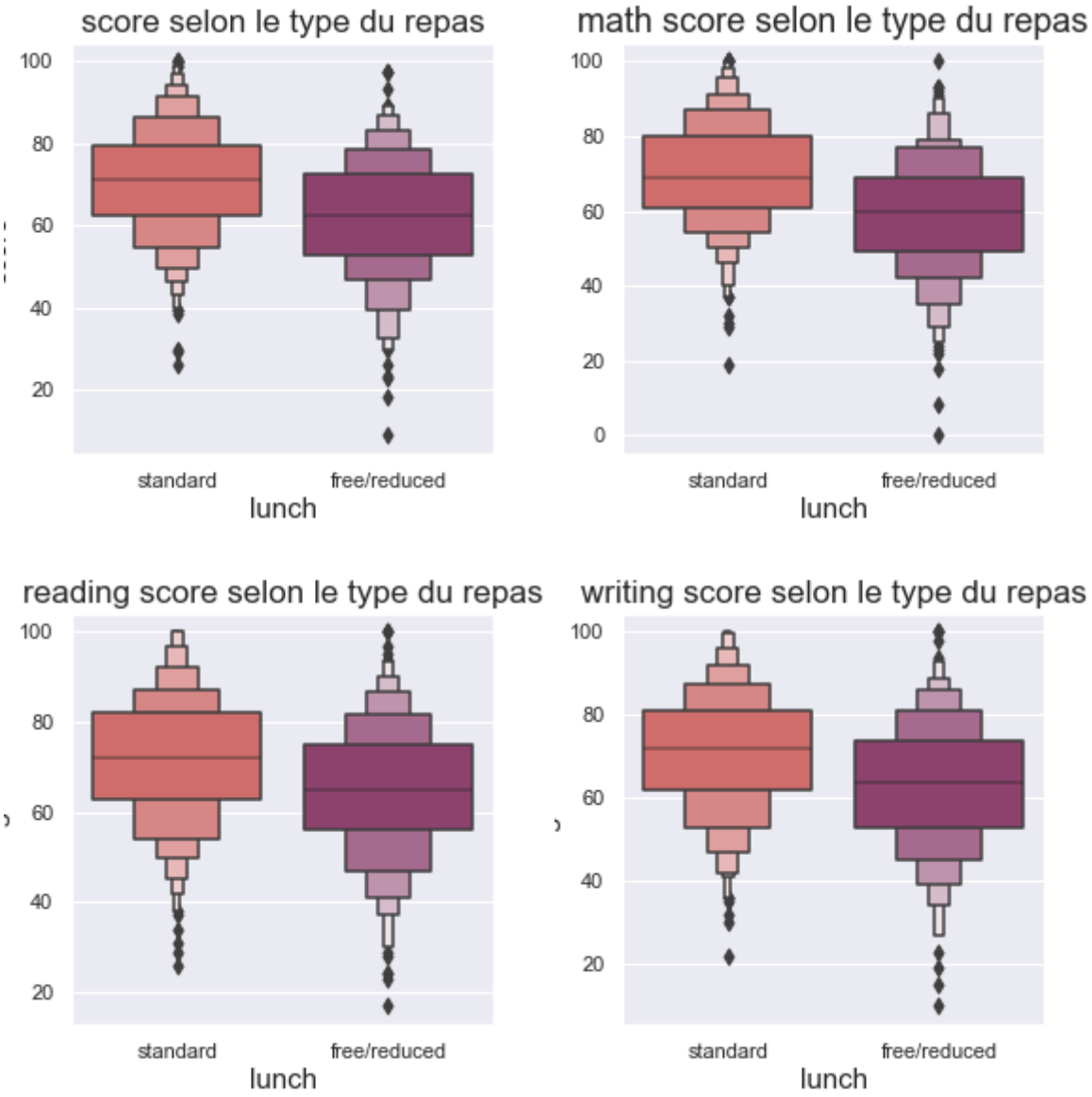




Bien que les hommes ont des scores en math supérieurs à 40%, et une médiane de 70%,dépassant celle des femmes ,ces dernières obtiennent de meilleurs résultats en écrit et en lecture, qui se répartissent sur une intervalle de [35 à 100] et des médianes de 70% .

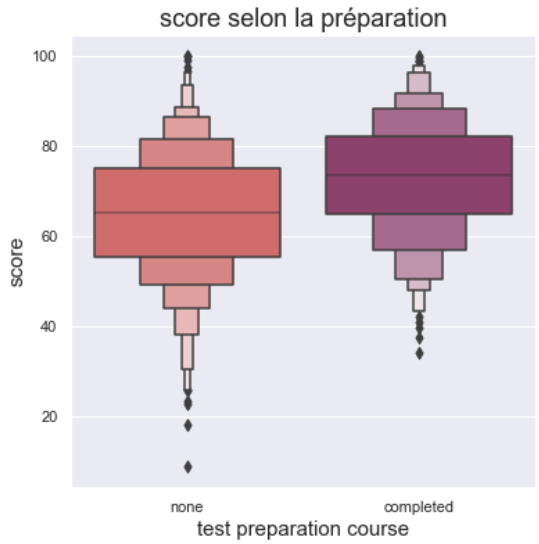
On suppose que le genre est un facteur influent sur le score .

## 2.2 Analyse des résultats selon le type de repas




Les scores des élèves ayant bénéficiés d'un repas standard atteignent les 100% et surpassent avec une médiane de 70% ceux des élèves qui prennent des repas réduits , et cette différence s'accentue surtout en mathématique .

## 2.3 Analyser les résultats selon leurs préparation à l'examen

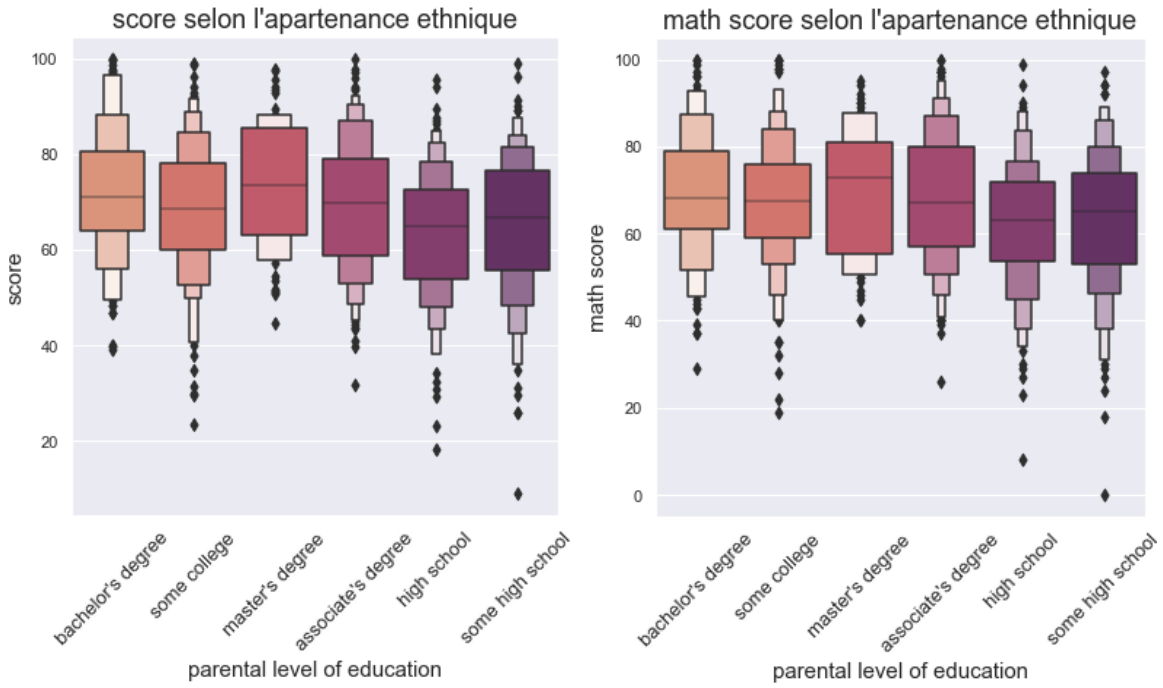






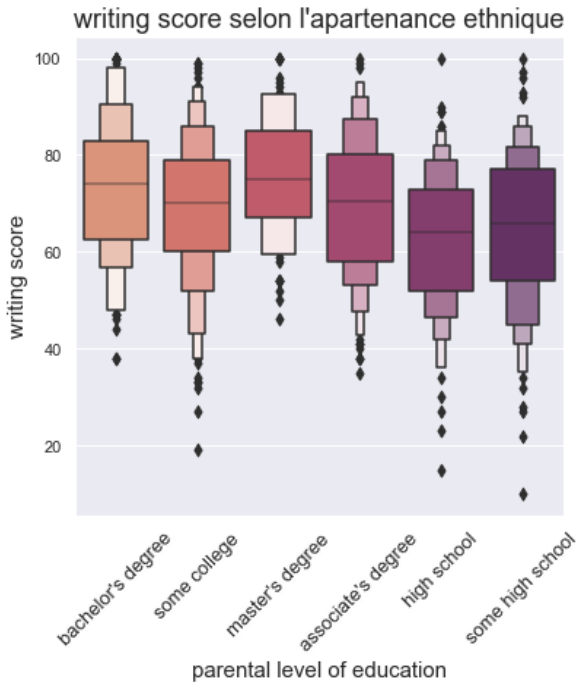
Les élèves ayant complété leurs cours de préparation aux examens, dépassent tous les 40% (à l'exception des outliers) et la médiane est de  et 75%, pendant que la médiane  de ceux qui ne se sont pas préparés  est de 65% pour diminuer jusqu'à 25%.

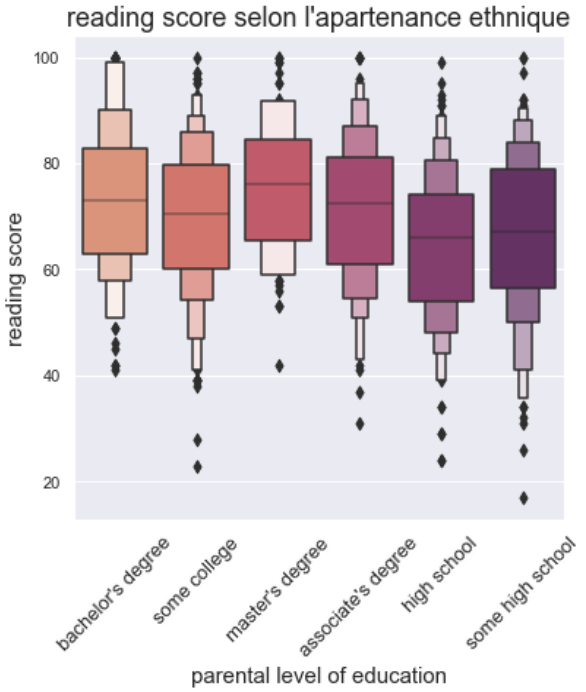
## 2.4 Analyser les résultats selon le niveau d'éducation des parents





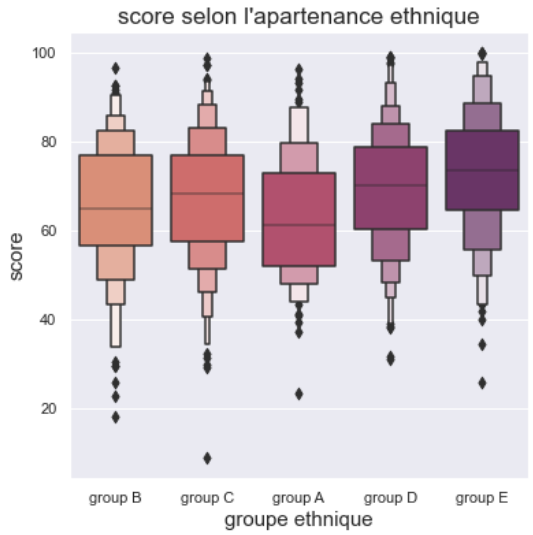


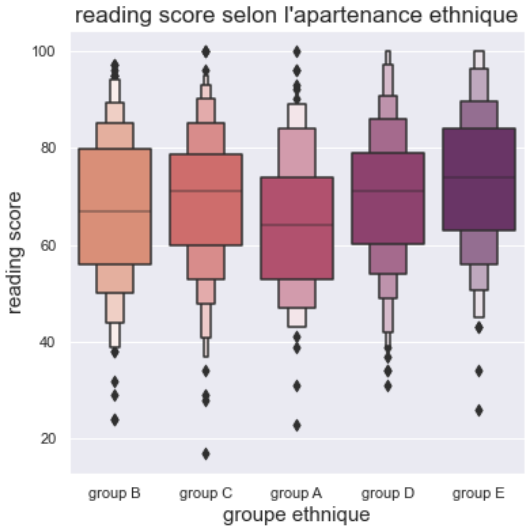


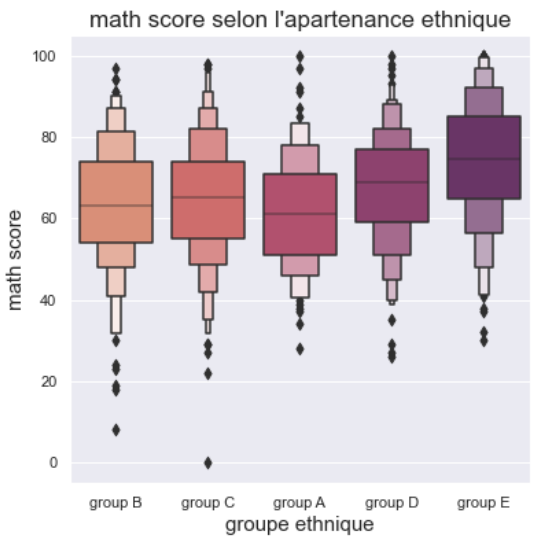



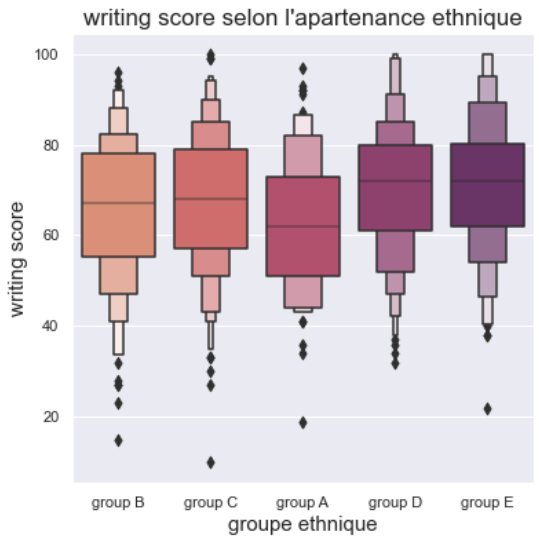
Les scores des élèves suivent la même logique de distribution que celle du niveau d'éducation de leurs parents , avec 75% de médiane et min dépassant les 60% pour la catégorie master, la médiane et l'intervalle minimum continue de baisser.

## 2.5 Analyse des résultats selon leurs appartenance ethnique:



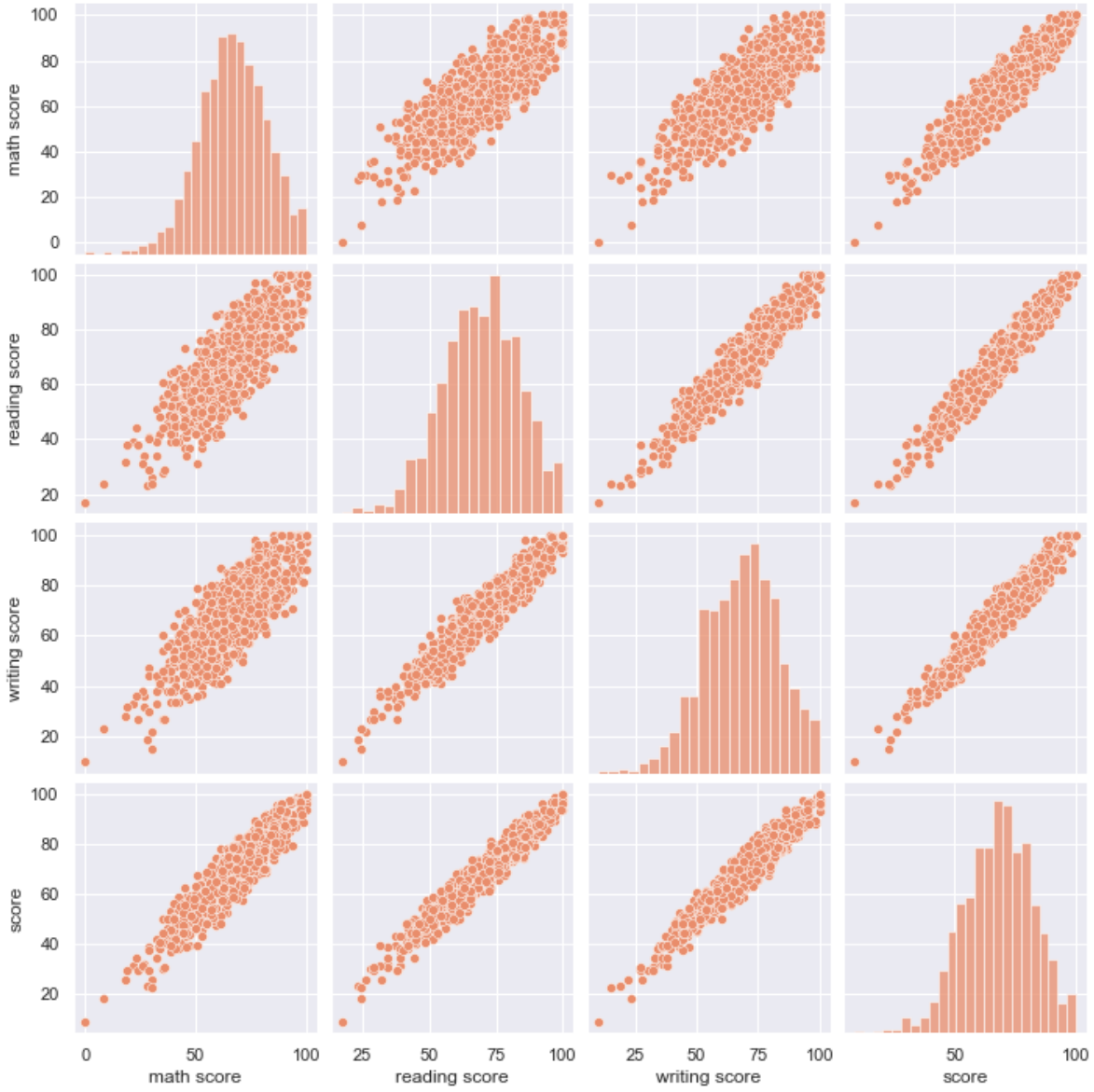




- Les groupes préservent leurs classements pour chacune des disciplines
- Le groupe E obtiennent les meilleures scores, surtout en math, ou (50% ont entre [65, 85], une médiane de 75% et un min de 40%),
- le groupe D l'église en writing.

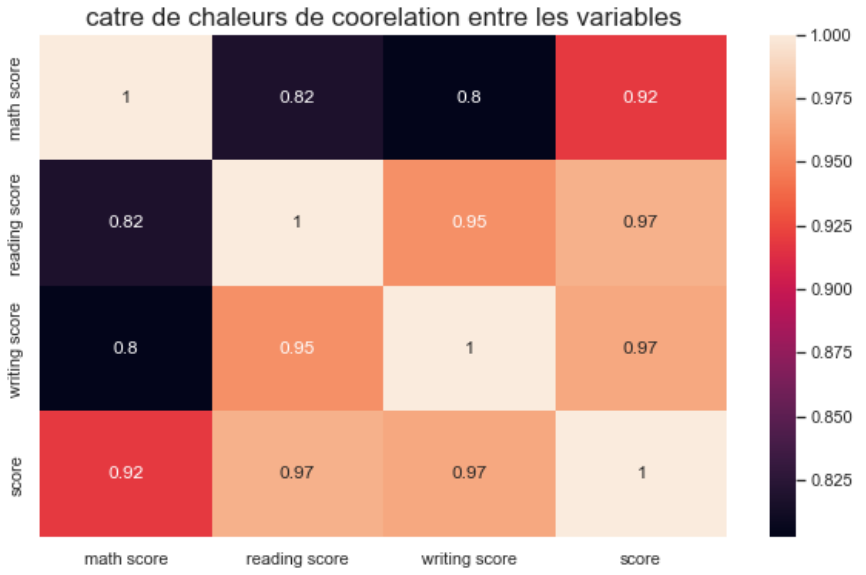
- Contrairement au groupeA  qui  représente les scores les plus faibles, ou 50% est compris entre 50 à 70 ].
- pour les groupes C & D, une légère différence avec de meilleurs résultats en reading et en writing qu'on math.

## 2.6 Analyse bi variée des scores des scores




- Les résultats tracent des droite linéaire allant toutes dans la même direction
- Il existe une  corrélation linéaire entre les résultats.

## 2.7 La corrélation entre les scores




- La carte de chaleur montre que les variables (scores) représentent un degré élevé,
- Les variables sont fortement corrélées.

# 3 Analyser les facteurs influençant les résultats

En appliquant une régression linéaire sur les données pour chacun des scores, j'ai pu confirmer mes hypothèses.

j'ai tenté d'expliquer les scores obtenus,à partir des variables **exogènes**(**explicatives**) [[niveau des parents, appartenance ethnique, préparation aux tests, le type du repas ainsi que le genre ], pour confirmer ou infirmer leurs taux et sens d' influence s' il existe.

NB:

- **gender**: 0:mal,1:femmel.
- **les niveaux d'éducation des parents**: sont classées du plus haut au plus bas pour le niveau des parents ce qui explique les coefficients négatifs:

  'master\'s degree':'1','bachelor\'s degree':'2', 'associate\'s degree':'3', 'high school':'4', 'some high school':'5', 'some college':'6'

- **les groupes:**

  'group A':'1',  'group B':'2',  'group C':'3', 'group D':'4',    'group E':'5'

- **Lunch:**'standard':'1','free/reduced':'0'

## 3.1 Régression linéaire pour le Score en math

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | math_score | R-squared: | 0.255 |
| Model: | OLS | Adj. R-squared: | 0.246 |
| Method: | Least Squares | F-statistic: | 28.12 |
| Date: | Sun, 31 Jul 2022 | Prob (F-statistic): | 2.62e-55 |
| Time: | 00:46:17 | Log-Likelihood: | -3990.3 |
| No. Observations: | 1000 | AIC: | 8007. |
| Df Residuals: | 987 | BIC: | 8070. |
| Df Model: | 12 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 60.0195 | 2.323 | 25.832 | 0.000 | 55.460 | 64.579 |
| gender[T.1] | -4.9953 | 0.839 | -5.954 | 0.000 | -6.642 | -3.349 |
| ethnicity[T.2] | 2.0408 | 1.700 | 1.201 | 0.230 | -1.295 | 5.376 |
| ethnicity[T.3] | 2.4700 | 1.592 | 1.552 | 0.121 | -0.654 | 5.594 |
| ethnicity[T.4] | 5.3410 | 1.624 | 3.289 | 0.001 | 2.154 | 8.528 |
| ethnicity[T.5] | 10.1347 | 1.802 | 5.626 | 0.000 | 6.599 | 13.670 |
| parental_level_of_education[T.2] | -0.9223 | 2.107 | -0.438 | 0.662 | -5.058 | 3.213 |
| parental_level_of_education[T.3] | -2.8884 | 1.938 | -1.490 | 0.136 | -6.692 | 0.915 |
| parental_level_of_education[T.4] | -3.4710 | 1.930 | -1.798 | 0.072 | -7.259 | 0.317 |
| parental_level_of_education[T.5] | -7.6911 | 1.971 | -3.903 | 0.000 | -11.558 | -3.824 |
| parental_level_of_education[T.6] | -7.1371 | 1.990 | -3.586 | 0.000 | -11.043 | -3.231 |
| lunch[T.1] | 10.8768 | 0.873 | 12.463 | 0.000 | 9.164 | 12.589 |
| test_preparation_course[T.1] | 5.4947 | 0.876 | 6.275 | 0.000 | 3.776 | 7.213 |

| | | | |
|---|---|---|---|
| Omnibus: | 9.026 | Durbin-Watson: | 2.043 |
| Prob(Omnibus): | 0.011 | Jarque-Bera (JB): | 9.201 |
| Skew: | -0.232 | Prob(JB): | 0.0100 |
| Kurtosis: | 2.928 | Cond. No. | 16.2 |

## 3.2 Régression linéaire pour score en writing

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | writing_score | R-squared: | 0.334 |
| Model: | OLS | Adj. R-squared: | 0.326 |
| Method: | Least Squares | F-statistic: | 41.25 |
| Date: | Sun, 31 Jul 2022 | Prob (F-statistic): | 8.17e-79 |
| Time: | 00:53:01 | Log-Likelihood: | -3936.2 |
| No. Observations: | 1000 | AIC: | 7898. |
| Df Residuals: | 987 | BIC: | 7962. |
| Df Model: | 12 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 57.9422 | 2.201 | 26.323 | 0.000 | 53.623 | 62.262 |
| gender[T.1] | 9.0965 | 0.795 | 11.444 | 0.000 | 7.537 | 10.656 |
| ethnicity[T.2] | 1.2201 | 1.610 | 0.758 | 0.449 | -1.940 | 4.380 |
| ethnicity[T.3] | 2.4126 | 1.508 | 1.600 | 0.110 | -0.547 | 5.372 |
| ethnicity[T.4] | 5.9307 | 1.539 | 3.855 | 0.000 | 2.911 | 8.950 |
| ethnicity[T.5] | 5.1373 | 1.707 | 3.010 | 0.003 | 1.788 | 8.487 |
| parental_level_of_education[T.2] | -1.6983 | 1.997 | -0.851 | 0.395 | -5.616 | 2.220 |
| parental_level_of_education[T.3] | -5.1832 | 1.836 | -2.823 | 0.005 | -8.787 | -1.580 |
| parental_level_of_education[T.4] | -6.1037 | 1.829 | -3.338 | 0.001 | -9.692 | -2.515 |
| parental_level_of_education[T.5] | -10.9976 | 1.867 | -5.891 | 0.000 | -14.661 | -7.334 |
| parental_level_of_education[T.6] | -10.5054 | 1.886 | -5.571 | 0.000 | -14.206 | -6.805 |
| lunch[T.1] | 8.2028 | 0.827 | 9.921 | 0.000 | 6.580 | 9.825 |
| test_preparation_course[T.1] | 10.0587 | 0.830 | 12.125 | 0.000 | 8.431 | 11.687 |

| | | | |
|---|---|---|---|
| Omnibus: | 16.647 | Durbin-Watson: | 2.038 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 17.222 |
| Skew: | -0.321 | Prob(JB): | 0.000182 |
| Kurtosis: | 2.982 | Cond. No. | 16.2 |

## 3.3 Régression linéaire pour score en reading

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | reading_score | R-squared: | 0.227 |
| Model: | OLS | Adj. R-squared: | 0.218 |
| Method: | Least Squares | F-statistic: | 24.19 |
| Date: | Sun, 31 Jul 2022 | Prob (F-statistic): | 8.62e-48 |
| Time: | 00:51:09 | Log-Likelihood: | -3970.5 |
| No. Observations: | 1000 | AIC: | 7967. |
| Df Residuals: | 987 | BIC: | 8031. |
| Df Model: | 12 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 61.0488 | 2.278 | 26.798 | 0.000 | 56.578 | 65.519 |
| gender[T.1] | 7.0714 | 0.823 | 8.596 | 0.000 | 5.457 | 8.686 |
| ethnicity[T.2] | 1.3260 | 1.667 | 0.796 | 0.426 | -1.944 | 4.596 |
| ethnicity[T.3] | 2.2739 | 1.561 | 1.457 | 0.145 | -0.789 | 5.337 |
| ethnicity[T.4] | 4.1056 | 1.592 | 2.578 | 0.010 | 0.981 | 7.230 |
| ethnicity[T.5] | 5.5135 | 1.766 | 3.121 | 0.002 | 2.047 | 8.980 |
| parental_level_of_education[T.2] | -2.0491 | 2.066 | -0.992 | 0.322 | -6.104 | 2.006 |
| parental_level_of_education[T.3] | -4.2051 | 1.900 | -2.213 | 0.027 | -7.934 | -0.476 |
| parental_level_of_education[T.4] | -5.4846 | 1.893 | -2.898 | 0.004 | -9.199 | -1.770 |
| parental_level_of_education[T.5] | -9.1055 | 1.932 | -4.713 | 0.000 | -12.897 | -5.314 |
| parental_level_of_education[T.6] | -8.2541 | 1.951 | -4.230 | 0.000 | -12.084 | -4.425 |
| lunch[T.1] | 7.2458 | 0.856 | 8.468 | 0.000 | 5.567 | 8.925 |
| test_preparation_course[T.1] | 7.3625 | 0.859 | 8.576 | 0.000 | 5.678 | 9.047 |

Le $R^2$ <30% peuvent remettre en cause sa  performance de ces modèles .

Néanmoins les coefficients nous démontrent combien certaines catégories des variables explicatives influent sur les résultats.

**Gender:**

**NB:**pour cette variable les paramètres sont significatifs avec des (p-value 0.0<5%)

- score en math:pour les femmes , le coefficient diminue de - 5%, contrairement en lecture ou il augmente de 9% et de 7% en écriture .

**Ethnicity**:

**NB:**pour cette variable les paramètres significatifs sont ceux des groupes D et E avec des (p-value<5%)

- le groupe E,influence sur les scores de (5 à 6 %) de plus en math et en ecrit par rapport au groupe A .
- le groupe D de même que le E influence positivement avec un coefficient 5% en math et 5% en écrit 4% en lecture.

**Le niveau d'éducation des parents:**

- Plus on descend en niveau, les coefficients s'accroissent dans le sens négatif ,uniquement pour les niveaux high school et somme high school ou ils augmentent modestement.

**Les cours de préparation à l'examen:**

- les élèves ayant accompli leurs cours de préparations à l'examen ont un score de de 5,5% plus que les autres (paramètre significative ;p-value >5%)

**le type de repas:**

- Vu la significativité du paramètre( 0.00 < 5%) on peut confirmer que le repas complet participe de 11% à remonter le score en math et de 8% en reading et 7% en writing par rapport au repas réduit.

## 3.4 Synthèse

On peut conclure que malgré le $R^2$ non performant <50% ,on arrive à expliquer 30% les résultats et à démontrer l'influence de ces variables exogènes tel que:

- ❖ Les hommes ont un meilleur score en math que les femmes , à l'inverse en lecture et l'écrit ou ces dernières sont en tête.
- ❖ on obtient de meilleurs résultats avec la catégorie des élèves ayant des parents avec un meilleur niveau d'éducation.
- ❖ Les repas , quand ils sont complets , contribuent à la performance des scores.

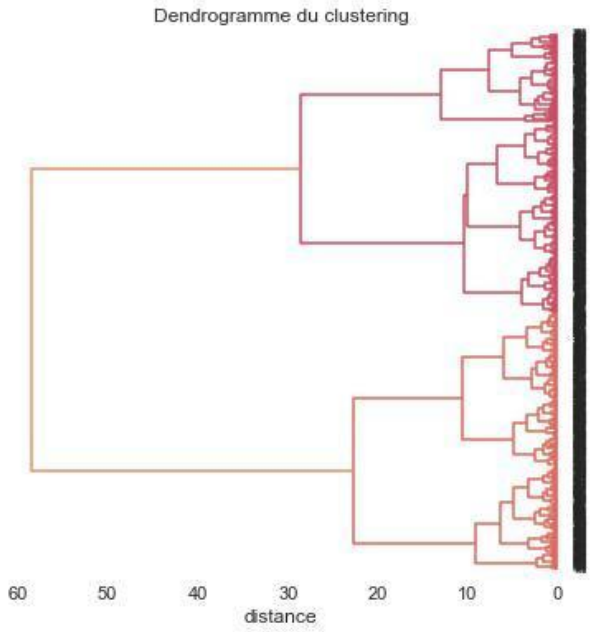
Il reste 70% qu'on arrive pas à expliquer avec ces facteurs .

# 4 Analyse exploratoire des données

Pour mener à bien mes analyses il est opportun de classer les élèves selon leurs résultats, pour ce faire on utilise une classification non supervisée.

## 4.1dendrogramme de classification hiérarchique

Pour avoir une idée sur les classifications possibles et les groupes qu'ils peuvent former, on construit un dendrogramme de classification hiérarchique avec des données brutes .

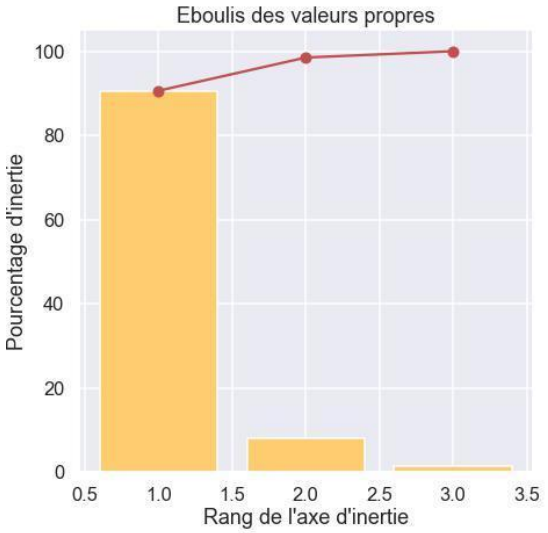


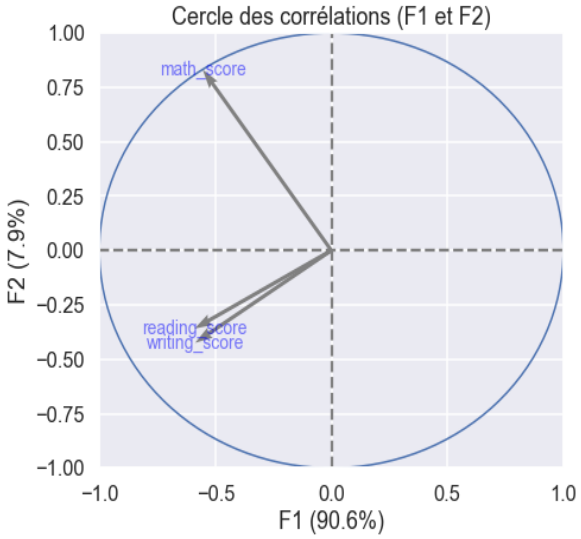
En observant la répartition des classes, on peut opter pour la classification en 2 ou 4 sous ensembles formant des groupes équilibrés

## 4.2 Analyse en Composantes Principales

On commence par l'étudier de la liaisons entre les variables, et les regrouper en de nouvelles variables synthétiques grâce à une **ACP**, en gardant le maximum d'informations.

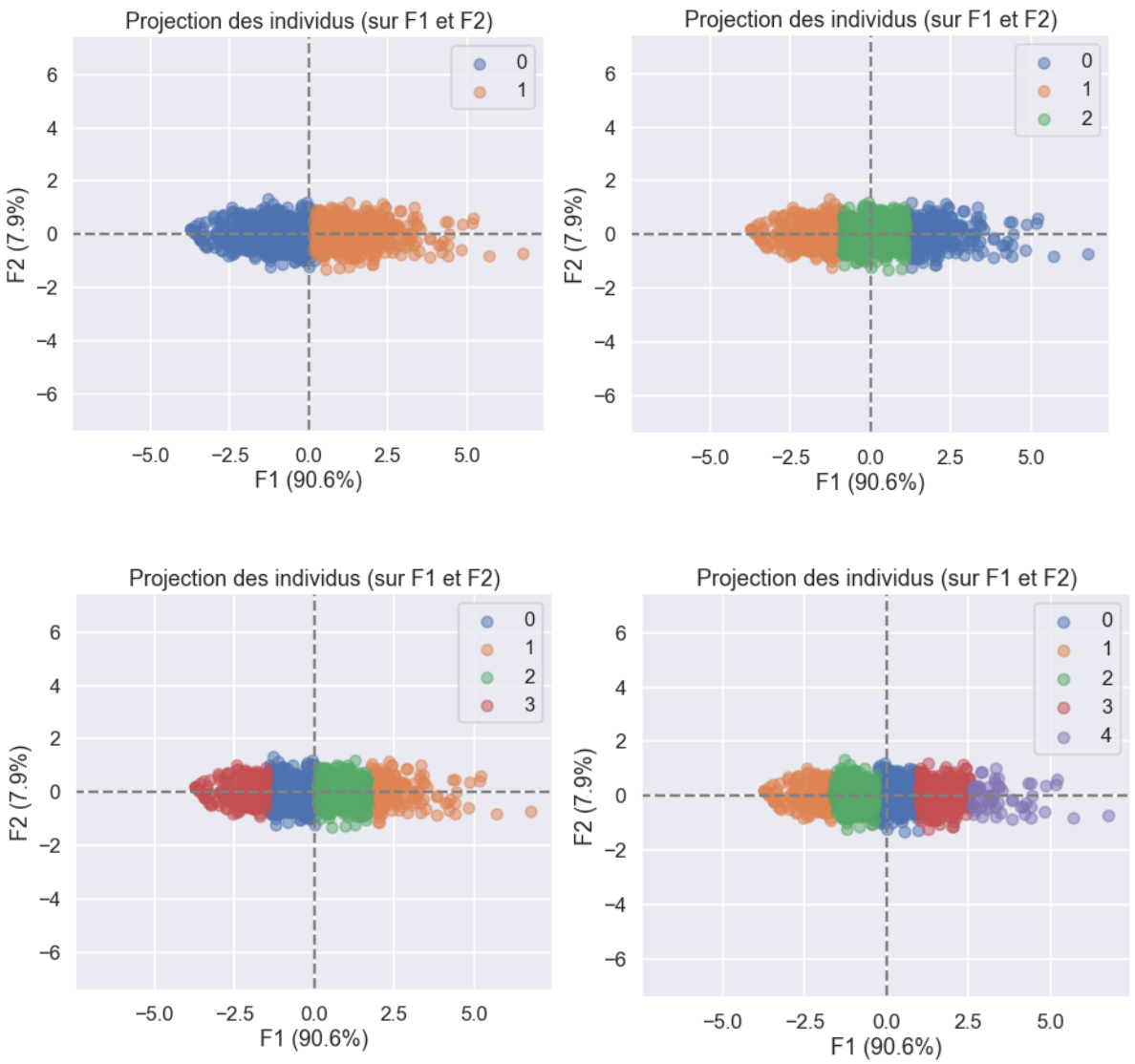
Pour ce faire, trois composantes sont prises en compte (**math scores, reading score, writing score).**




- l'aboulie des valeurs propres démontre qu'un seul plan est nécessaire pour la projection de nos individus (90 % des variables se projettent sur le premier plan factoriel),
- On trace un cercle de corrélation, ou on peut lire que les variables sont corrélées à F1 (dans le sens négatif )et à F2 (avec var math dans le sens positif et reading et writing de dans le sens négatif) .

## 4.3 Clustering avec le k-means

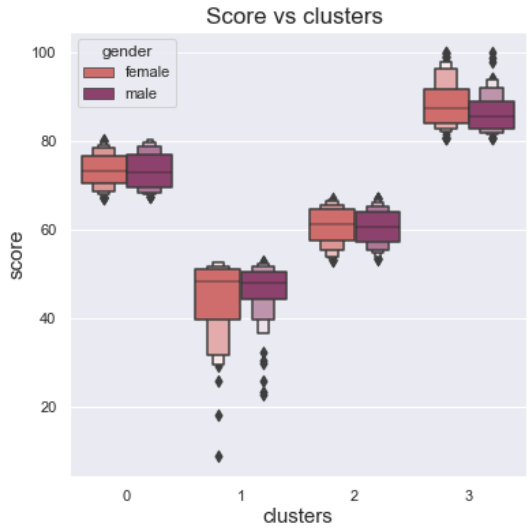



Le partitionnement en 4 clusters semble la plus adapté, et forme des groupes presque de même taille qui s'étalent sur l'axe F1 représentant la moyenne des scores. Avec une distribution symétrique au sein d'un même groupe :

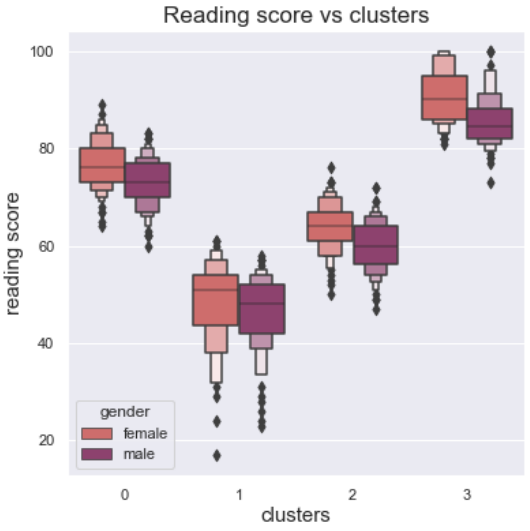
- **Le cluster 3 et 0** se situent dans le sens positif de F2(élèves les plus forts).
- **Le cluster 2 et 1** se situent dans le sens négatif fe F2.
- On peut détecter l'orientation des élèves de chaque groupe ;plus littéraires(élèves situés dans la partie négative de F2 , ou plus scientifiques pour ceux situés dans la partie positif de F2)
- mais dans le cas général, les élèves ont presque le même niveau dans les deux disciplines (ce qui explique la linéarité des variables).

## 4.4 Analyser les clusters

Dans cette partie des analyses multivariées ont été effectuées sur les clusters avec chacune des variables



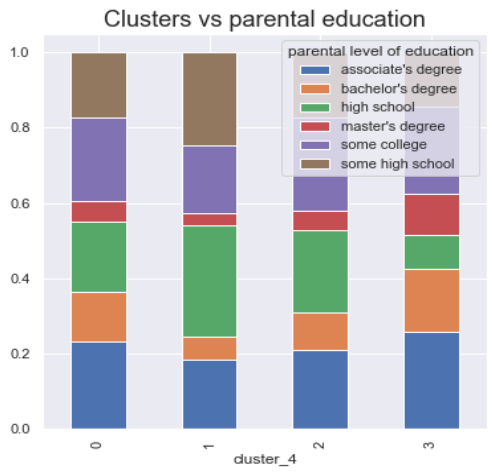
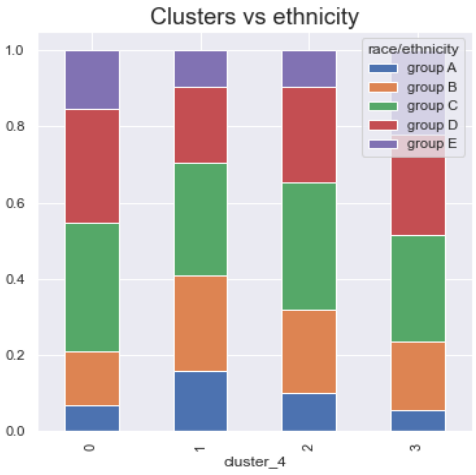


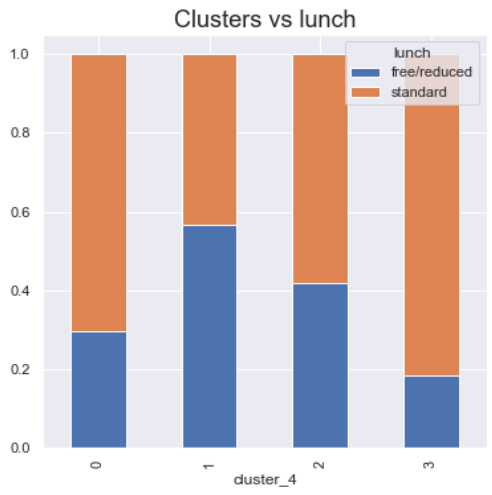
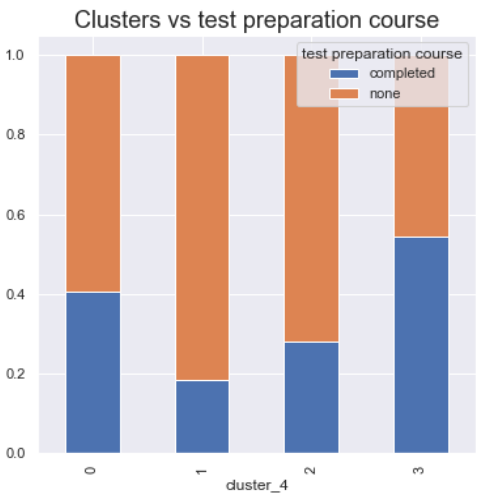




Les clusters pour chacun des modules, suivent un même classement selon les scores, avec un léger décalage entre les femmes et les hommes ;

Les trois clusters dépassent la moyenne, avec le cluster 3 qui est en tête du classement (une médiane de 90% et une intervalle comprise entre 80 - 100). succédé par le cluster0,(score[70,80])(score[70,80],puis le cluster 2 avec un score de [50-70].Enfin le cluster1 regroupant les élève au dessous de la moyenne, avec un score de [30-50]

Dans tous les clusters, les femmes ont de meilleurs scores en lecture et en écrit, mais arrivent deuxièmes en math, ce qui fait qu'en moyenne ils sont égaux.

<table>
<tr>
<td> 
</td>
<td>
<ul>
<li>le cluster 3 (le plus fort)ont une plus grande part des parents avec un niveau BAC+2 par rapport aux autres groupes.</li>
<li>vient ensuite le cluster 0</li>
<li>puis le 2 et le 1 successivement avec moins de parents en master et bachelor et plus en high et somme école.</li>
</ul>
Le classement des clusters dépend en partie du niveau des parents.
</td>
</tr>
<tr>
<td>
<ul>
<li>Les individus du groupe E et D se concentrent dans les clusters 3 puis 0 ,</li>
<li>se rétrécissent dans le cluster 2, surtout le 1 (le plus faible) pour permettre aux groupes A,B et C de s'accroître .</li>
</ul>
la part des groupe ethnique, facteur déterminant dans le classement des clusters
</td>
<td> 
</td>
</tr>
</table>

| | |
|---|---|
|  |  |
| La part des élèves ayant bénéficiés d'un repas complet est très importante dans les clusters 3 et 0, et diminue parallèlement au classement des clusters. | Les clusters les mieux classées comptent une plus grande part des élèves les mieux préparés à l'examen. |

# Conclusion

Parmi les facteurs influençant l'éducation et la réussite des élèves :

- Le genre a une modeste influence sur l’orientation des élèves.
- Les conditions socio-culturelles (appartenance ethnique et niveau d’éducation des parents ,
- Les conditions économiques (type du repas ).
- la motivation et la préparation des élèves.

Avec un peu de préparation et de motivation les élèves peuvent réussir dans toutes les disciplines.

L'éducation permet non seulement de former les générations présentes mais aussi les générations futures.

L'inégalité contribue à l'échec scolaire .

En leur offrant de meilleures conditions, on peut remonter le niveau des élèves.

Il existe d’autres facteurs qui influencent les résultats des élèves, reste un sujet intéressant à explorer.