Rapport DUT STID

Année scolaire : 2020/2021

Étude analytique sur la performance des étudiants pendant les examens

Par :

Sâm RAHIMI

Lukasz GUMIENIAK

Don Melin Lorensky THERTULIEN

Encadrée par :
M. Omar OSMANI

INTRODUCTION

L'éducation est l'un des piliers majeurs indispensable au développement économique, social, culturel d'une société. En effet, elle aide au changement, à l'amélioration des conditions de vies des populations du monde entier. Dans le cadre d'un projet d'analyse au cours de notre 4e semestre de formation STID, nous avons donc décidé de réaliser une étude en lien avec l'éducation **: la performance des étudiants durant les examens.**

En nous intéressant sur ce thème des examens, nous avons voulu faire le tour sur les différents aspects qui peuvent aider à comprendre et ainsi améliorer les systèmes d'éducation actuels.

Dans cette optique, nous avons approcher le côté social et culturel de notre sujet grâce à une base de données et différents outils de programmation/ et ou visualisation.

Notre étude a débuté au début du mois de février et s'est déroulée sur deux mois.

Problématiques :

Nous nous sommes posés les questions suivantes pour tenter de nous rapprocher le plus de notre but :

- Quelles sont les matières les plus réussies ?
- Quel genre réussi le mieux les examens ?
- Les résultats sont-ils corrélés avec le niveau d'étude des parents ?
- Est-ce que la préparation avant les examens à une influence sur les notes obtenues ?
- Le niveau d'étude des parents est-il en lien avec la préparation des étudiants au vu des examens ?

ÉTAT DE L'ART

Notre analyse se penchera donc sur différents axes tels que les matières les plus réussies ou bien le sexe qui réussit le mieux les examens ou encore l'impact des diplômes des parents.

Nous allons alors vous présenter tout d'abord, toutes les études que l'on a pu trouver qui ont été menées sur ce sujet.

Pour commencer, nous avons trouvé des travaux qui cherchaient à prédire la performance des étudiants pour différentes raisons : les universités veulent détecter les étudiants à risques, prévoir les ressources nécessaires pour les cours, surtout les universités qui visent à améliorer leur réputation, elles pourront ainsi connaître les étudiants qui se distinguent et surtout dans quels domaines ils excellent pour les affecter à certains problèmes réels qui sont à échelle mondiale.

Les recherches dont on va vous parler se basaient sur l'université Britannique de Dubaï. Les chercheurs ont voulu explorer une possible utilisation de la petite taille de l'ensemble de données des élèves car les dossiers des étudiants de cette université sont de petites tailles puisqu'elle est récente (2004).

En utilisant des algorithmes de visualisation et de clustering, Cette recherche explore également la possibilité d'identifier les indicateurs clés dans le petit ensemble de données, qui seront utilisés pour créer le modèle de prédiction

Ils ont utilisé les outils de programmation tels que Microsoft Excel, Python et Rstudio pour la visualisation des données.

Pour mener à bien leurs recherches, ils ont suivi ces 3 phases :

Phase de prétraitement de l'ensemble de données

Phase de nettoyage des ensembles de données

Phase de l'encodage des fonctionnalités

Leur jeu de données est constitué de 50 cas (dossiers des étudiants) et des 5 variables suivantes :

L'âge, le nom du baccalauréat, la note cumulée du baccalauréat, les cours suivis pendant leurs études de maîtrise avec leurs notes et le nom des instructeurs de chaque cours.

Ils se sont posés les questions suivantes :

- Quel est le meilleur modèle de classification par apprentissage automatique pour classer les notes des étudiants, en utilisant un ensemble de données de petite taille, avec un taux de précision raisonnable et significatif ?

- Quels sont les principaux indicateurs clés qui pourraient aider à créer le modèle de classification

- Les performances des étudiants dans n'importe quel cours pourraient-elles être prédites avec un taux de précision raisonnable et significatif en utilisant uniquement les dossiers de préadmission des étudiants, les noms de cours et les attributs du nom des instructeurs ?

Ils arrivent à conclure que les résultats prouvent la possibilité de le faire avec des taux d'exactitude raisonnablement significatifs

Aspect Business

Pour répondre à la question à savoir à qui cette étude pourrait être intéressante, la première pensée parait évidente : l'État. Effectivement, si on met de côté son intérêt pour augmenter le taux de réussite de sa population par rapport aux autres écoles étrangères, nous pouvons comprendre sa volonté de mettre en œuvre des plans d'aide pour les étudiants afin de faciliter leur apprentissage aux vus des examens. On peut percevoir l'aide au logement (APL) ainsi que l'aide financière (la Bourse) qui permettent de soulager les finances pour par exemple des cours particuliers.

Les travaux de cette étude pourraient également intéresser des plateformes de cours à distance comme par exemple le célèbre site de cours en ligne OpenClassrooms. Ce dernier propose des cours certifiants sur différentes thématiques et matières de manière gratuite et il offre des parcours débouchant sur des métiers en croissance.

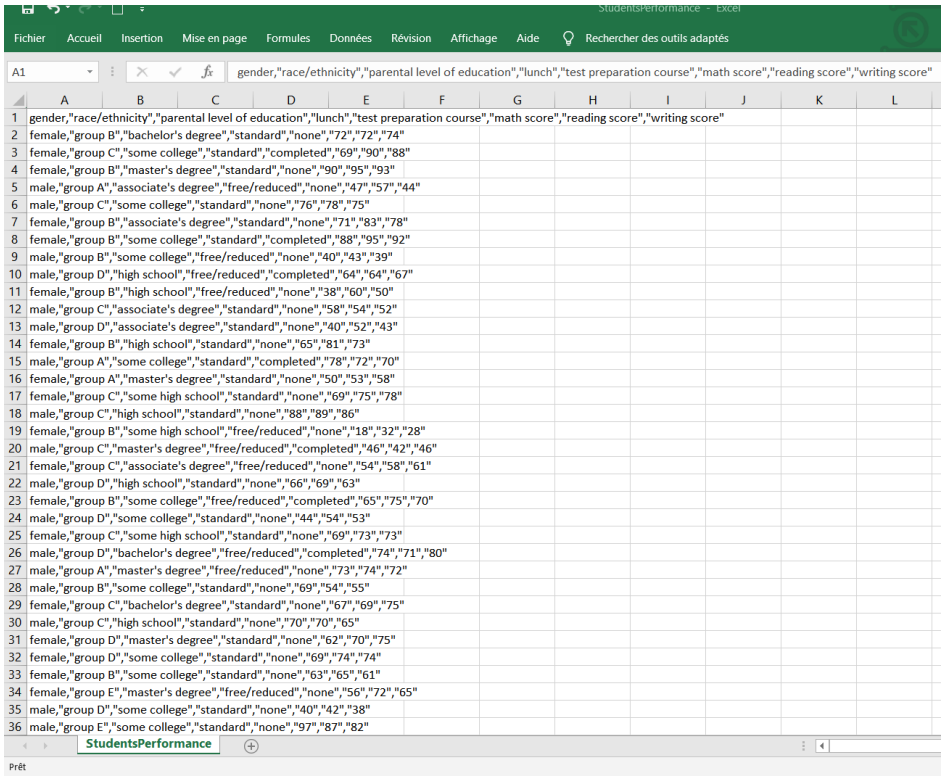
Il n'y a évidemment pas que les sites en ligne qui aimeraient profiter de ces analyses. Tout comme dit dans la partie *état de l'art*, les universités pourraient développer le potentiel maximum de leurs étudiant grâce aux résultats de notre étude. Cela leur permettrait de concurrencer les autres universités et ainsi, se faire une réputation à l'échelle mondiale tout en contribuant et encourageant le développement. On peut finalement citer encore les cours particuliers dans le cadre des soutiens scolaires.

Aspect technique

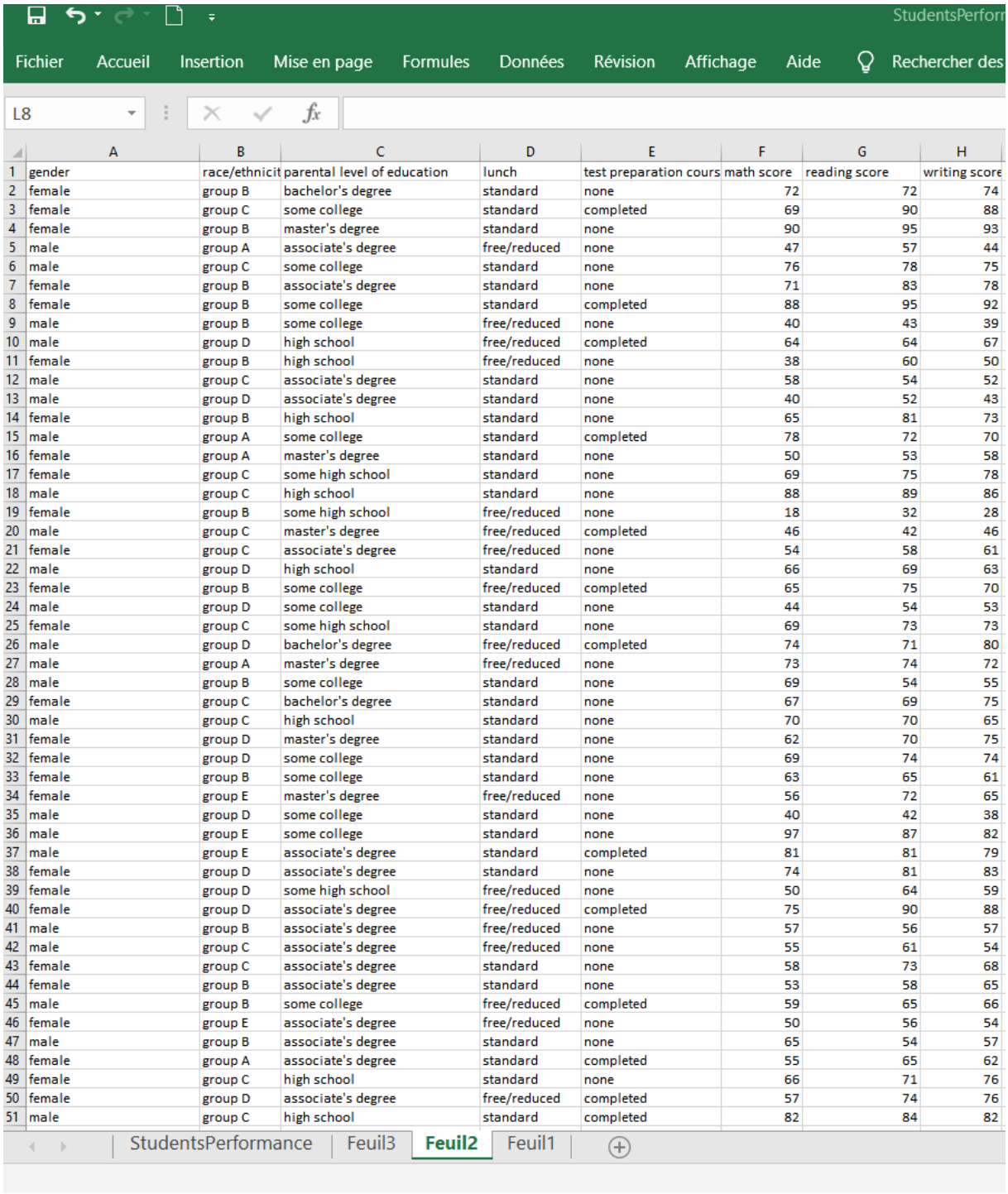
Notre jeu de données provient du site *Kaggle.* Il compte 1000 individus qui correspondent aux étudiants dont on a relevé 8 variables. Parmi ces variables, il y en a 5 qui sont qualitatives et les 3 autres sont quantitatives, nous avions dû modifier le fichier Excel de départ pour pouvoir l'exploiter au maximum.

Toutes les variables sont prises en compte pour nous aider à répondre aux questions que nous nous sommes posées, excepté "Race/Ethnicity"

Capture d'écran du jeu de données brute



Capture d'écran du jeu de données pour exploitation



StudentsPerform

Fichier Accueil Insertion Mise en page Formules Données Révision Affichage Aide Rechercher des

L8

	A	B	C	D	E	F	G	H
1	gender	race/ethnicit	parental level of education	lunch	test preparation cours	math score	reading score	writing score
2	female	group B	bachelor's degree	standard	none	72	72	74
3	female	group C	some college	standard	completed	69	90	88
4	female	group B	master's degree	standard	none	90	95	93
5	male	group A	associate's degree	free/reduced	none	47	57	44
6	male	group C	some college	standard	none	76	78	75
7	female	group B	associate's degree	standard	none	71	83	78
8	female	group B	some college	standard	completed	88	95	92
9	male	group B	some college	free/reduced	none	40	43	39
10	male	group D	high school	free/reduced	completed	64	64	67
11	female	group B	high school	free/reduced	none	38	60	50
12	male	group C	associate's degree	standard	none	58	54	52
13	male	group D	associate's degree	standard	none	40	52	43
14	female	group B	high school	standard	none	65	81	73
15	male	group A	some college	standard	completed	78	72	70
16	female	group A	master's degree	standard	none	50	53	58
17	female	group C	some high school	standard	none	69	75	78
18	male	group C	high school	standard	none	88	89	86
19	female	group B	some high school	free/reduced	none	18	32	28
20	male	group C	master's degree	free/reduced	completed	46	42	46
21	female	group C	associate's degree	free/reduced	none	54	58	61
22	male	group D	high school	standard	none	66	69	63
23	female	group B	some college	free/reduced	completed	65	75	70
24	male	group D	some college	standard	none	44	54	53
25	female	group C	some high school	standard	none	69	73	73
26	male	group D	bachelor's degree	free/reduced	completed	74	71	80
27	male	group A	master's degree	free/reduced	none	73	74	72
28	male	group B	some college	standard	none	69	54	55
29	female	group C	bachelor's degree	standard	none	67	69	75
30	male	group C	high school	standard	none	70	70	65
31	female	group D	master's degree	standard	none	62	70	75
32	female	group D	some college	standard	none	69	74	74
33	female	group B	some college	standard	none	63	65	61
34	female	group E	master's degree	free/reduced	none	56	72	65
35	male	group D	some college	standard	none	40	42	38
36	male	group E	some college	standard	none	97	87	82
37	male	group E	associate's degree	standard	completed	81	81	79
38	female	group D	associate's degree	standard	none	74	81	83
39	female	group D	some high school	free/reduced	none	50	64	59
40	female	group D	associate's degree	free/reduced	completed	75	90	88
41	male	group B	associate's degree	free/reduced	none	57	56	57
42	male	group C	associate's degree	free/reduced	none	55	61	54
43	female	group C	associate's degree	standard	none	58	73	68
44	female	group B	associate's degree	standard	none	53	58	65
45	male	group B	some college	free/reduced	completed	59	65	66
46	female	group E	associate's degree	free/reduced	none	50	56	54
47	male	group B	associate's degree	standard	none	65	54	57
48	female	group A	associate's degree	standard	completed	55	65	62
49	female	group C	high school	standard	none	66	71	76
50	female	group D	associate's degree	free/reduced	completed	57	74	76
51	male	group C	high school	standard	completed	82	84	82

StudentsPerformance | Feuil3 | Feuil2 | Feuil1

Outils et méthodologie

Nous allons à présent présenter les outils et la méthodologie qui vont nous permettre de mener l'étude sur la performance des étudiants lors des examens.

Tout d'abord, nous allons procéder au nettoyage des données brutes pour constater d'éventuelles variables qui nous seraient inutile. Nous effectuerons cette tâche à l'aide d'Excel et de Power Bi :

Sur ce logiciel, nous pourrons commencer à effectuer différents tableaux croisés dynamiques entre nos différentes variables, pour observer les premiers résultats susceptibles de nous orienter vers nos conclusions.

Ensuite, nous pensons à exporter le fichier Excel sur Power BI afin d'avoir une analyse et une visualisation plus globale du sujet.

Cela nous permettra de savoir sur quels axes s'appuiera notre étude et sur quels modèles se diriger.

Pour ce faire, nous allons prendre dans un premier temps le graphique de nuage de points afin de distinguer les données de chaque variable. Pour la première question posée, les variables utilisées seront les matières avec les étudiants tout genre confondu. Pour la deuxième question posée, nous utiliserons les variables des genres dans toutes les matières afin d'identifier les performances par genre. Enfin, pour la dernière question posée, les variables du degré d'étude des parents ainsi que la performance pour chaque étudiant seront utilisées.

Vient ensuite, la partie programmation, où à l'aide du langage Python, nous pourrons commencer notre analyse de manière plus détaillée pour obtenir des résultats qui répondraient aux questions posées précédemment.

Nous pensions tout d'abord à faire une ACP (analyse en composantes principales) pour confirmer d'éventuelles corrélations entre les variables qui nous intéresse, mais nous avons remarqué que cette méthode était plus utilisée lorsque le jeu de données comporte nombreux énormément d'individus et variables quantitatives, or le nôtre n'est composé que de 8 variables que l'on va réduire pour n'utiliser que celles qui nous sont utiles.

Ensuite, il n'y a que 3 variables quantitatives, nous nous sommes alors tournés vers une étude où nous pourront utiliser la corrélation de Pearson.

Résultats attendus

Pour commencer, nous pensons que les tests écrits et oraux seront mieux réussis que les tests en mathématiques, avec une meilleure réussite à l'écrit.

En ce qui concerne le genre qui aboutit à de meilleurs résultats, nous pouvons imaginer que les femmes réussissent mieux que les hommes du fait qu'elles sont plus studieuses et rigoureuses de manière générale.

Nous pouvons également présumer qu'il existe une certaine corrélation entre le niveau d'étude des parents par rapport aux résultats, mais aussi quant à la préparation des étudiants au vu des examens.

Pour finir notre analyse, on peut prédire que cette préparation a une influence sur les résultats.

Nos résultats seront présentés globalement sous forme de nuage de points, de diagrammes à bâton et de tableaux.

SYNTHESE ANALYTIQUE

Sous Excel

Les premières statistiques montrent que les tests de lecture et d'écriture sont plus réussis que le test de mathématiques. (Voir tableau ci-dessous).

Le test de lecture est plus réussi que le test d'écriture en général. Notre première hypothèse stipulant que les tests écrits et oraux seront mieux réussis que les tests en mathématiques est donc confirmée.

Moyenne de math score	Moyenne de reading score	Moyenne de writing score
66,089	69,169	68,054

Concernant le sexe qui obtient les meilleurs résultats durant les examens, nous pouvons apercevoir que notre hypothèse disant que les femmes réussissent mieux que les hommes est également validée. Par contre, les hommes sont nettement plus fort généralement en mathématiques avec une moyenne générale de 68.6 contre 63.7. (Voir tableau ci-dessous).

gender	Moyenne de math score	Moyenne de reading score	Moyenne de writing score	Moyenne générale
female	63,63320463	72,60810811	72,46718147	69,56949807
male	68,72821577	65,47302905	63,31120332	65,83748271
TOTAL	**66,1807102**	**69,04056858**	**67,88919239**	**67,70349039**

Ici encore, notre hypothèse de départ est confirmée. En effet, la différence observée entre les étudiants préparés et non-préparés avant les examens montre dans les trois matières, que les étudiants ayants complété les cours de préparation aux tests réussissent mieux que ceux qui ne l'ont pas fait. (Voir tableau ci-dessous).

test preparation course	Moyenne de math score	Moyenne de reading score	Moyenne de writing score	Moyenne générale
completed	69,69553073	73,89385475	74,41899441	72,66945996
none	64,07788162	66,53426791	64,5046729	65,03894081
TOTAL	**66,88670617**	**70,21406133**	**69,46183366**	**68,85420039**

Sur le tableau qui suit, il apparait clairement que plus le niveau d'étude des parents de l'étudiant est élevé, mieux sont ses moyennes (en jaune nous avons les parents qui se sont arrêtés au lycée et qui n'ont pas terminé leur études, en vert sont représentés ceux qui ont obtenu un diplôme postbac jusqu'au master). Hypothèse validée.

parental level of education	Moyenne de math score	Moyenne de reading score	Moyenne de writing score	Moyenne générale
high school	62,1377551	64,70408163	62,44897959	63,09693878
some high school	63,4972067	66,93854749	64,88826816	65,10800745
some college	67,12831858	69,46017699	68,84070796	68,47640118
associate's degree	67,88288288	70,92792793	69,8963964	69,56906907
bachelor's degree	69,38983051	73	73,38135593	71,92372881
master's degree	69,74576271	75,37288136	75,6779661	73,59887006
TOTAL	**66,63029275**	**70,06726923**	**69,18894569**	**68,62883589**

Par contre nous pensions que plus le niveau d'étude des parents est élevé, plus l'étudiant suivrait et compléterait les cours de préparation. Or, ce n'est pas ce que nous affirme le tableau ci-dessous. Au contraire, les étudiants dont les parents ont un haut niveau d'étude ne suivent pas souvent les cours de préparation (2% contre 4% pour les enfants de master, ainsi de suite).

Cette fois-ci, notre hypothèse n'est pas validée.

	test preparation course	
parental level of education	completed	none
associate's degree	8%	14%
bachelor's degree	5%	7%
high school	6%	14%
master's degree	2%	4%
some college	8%	15%
some high school	8%	10%

Conclusion

Ainsi s'achèvent nos analyses sur Excel, avec la plupart de nos hypothèses qui se sont confirmées (4/5).

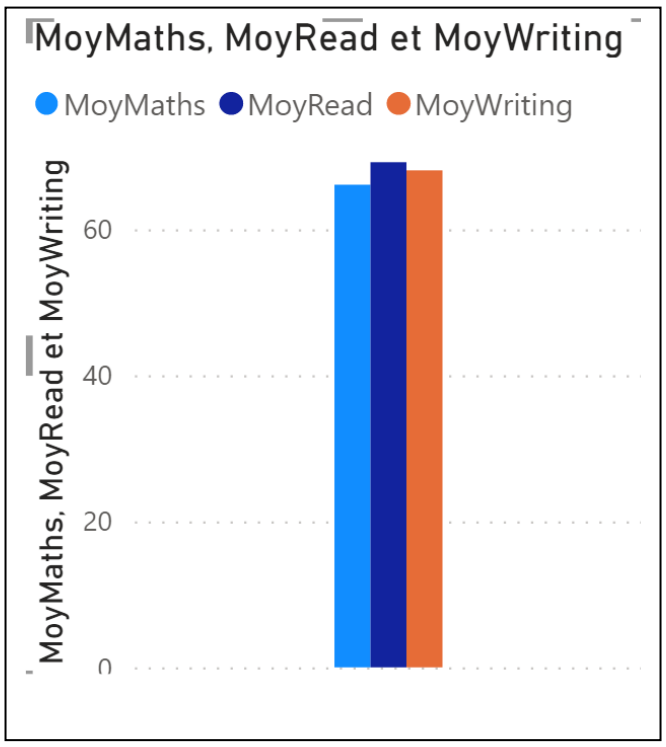
Sous Power Bi

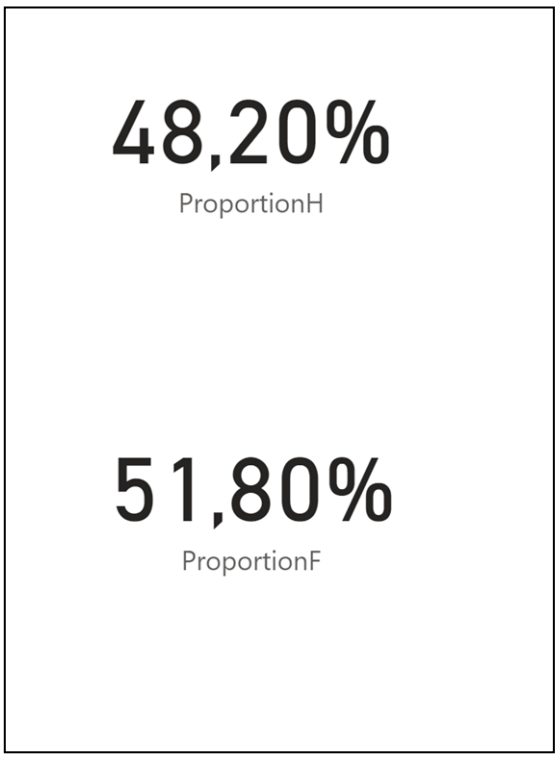
Nous allons maintenant présenter nos analyses visuelles faites grâce à l'outil Power Bi.

Il nous a fallu modifier et créer quelques variables pour obtenir les résultats suivants, mais nous n'allons pas rentrer dans ces détails.

Tout d'abord, nous avons le diagramme en bâton ci-dessous qui représente les moyennes des trois matières en concordance avec ce que nous avons expliqué auparavant, c'est-à-dire que les mathématiques (bleu clair) sont moins bien réussies que les deux autres.

A sa droite, nous avons la proportion d'homme et de femme qu'il y a dans le jeu de données : nous constatons qu'il y a quasiment la même proportion d'hommes que de femmes, ainsi nos résultats ne seront pas biaisés par rapport à l'équilibre du nombre d'hommes et de femmes.

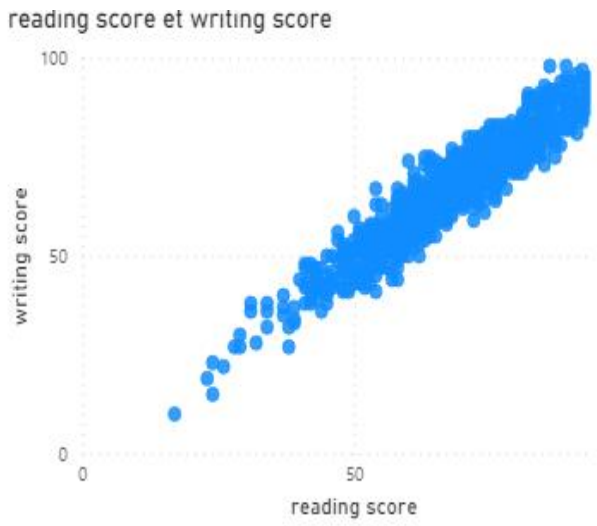


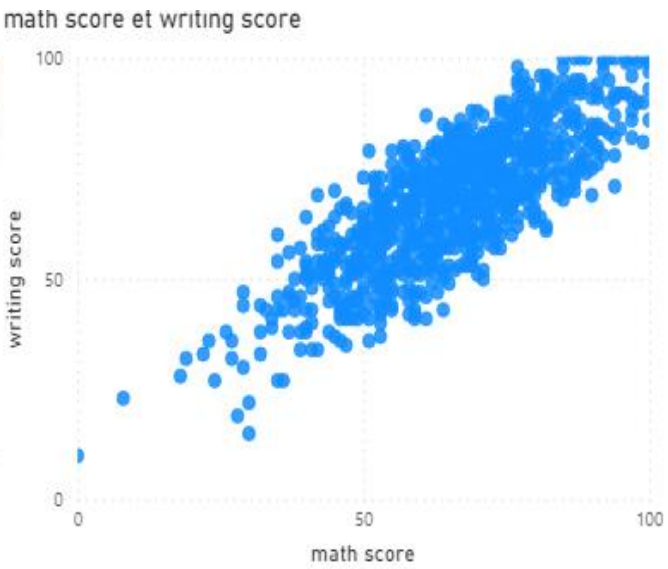


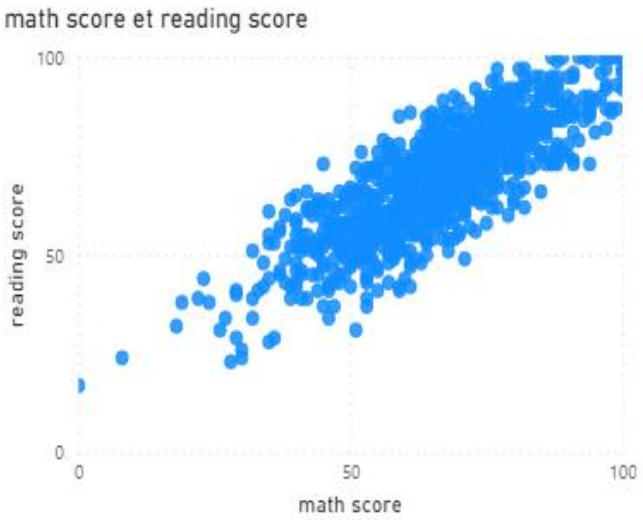
L’image ci-dessous correspond au croisement des trois matières entre elles grâce à des nuages de points.

Tout d’abord, il existe clairement une corrélation linéaire positive dans les trois cas.

Nous pouvons remarquer que les graphiques où est croisée la matière mathématiques avec les autres ont des points plus dispersés. Cela peut se traduire par un écart de réussite, en effet, nous avons précisé précédemment que les mathématiques étaient moins bien réussies que les autres.

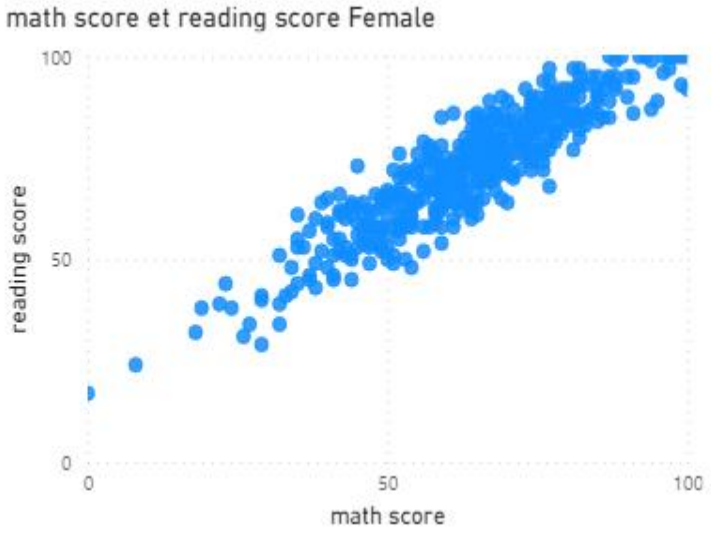


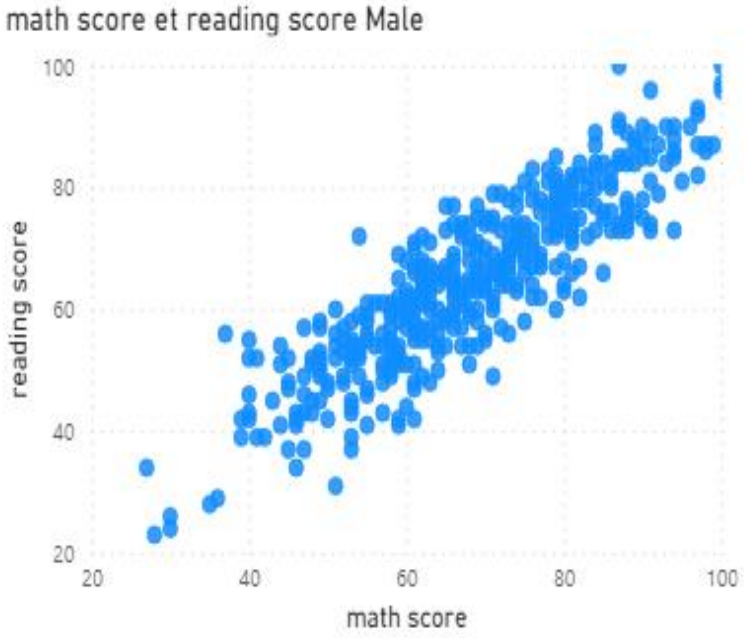




Ici, nous avons comparé les tests mathématiques et oraux des hommes et des femmes.

Le graphique représentant les hommes montre des points beaucoup plus dispersés que celui des femmes, nous pouvons interpréter ces résultats par la stabilité et la rigourosité des femmes en ce qui concerne les études.

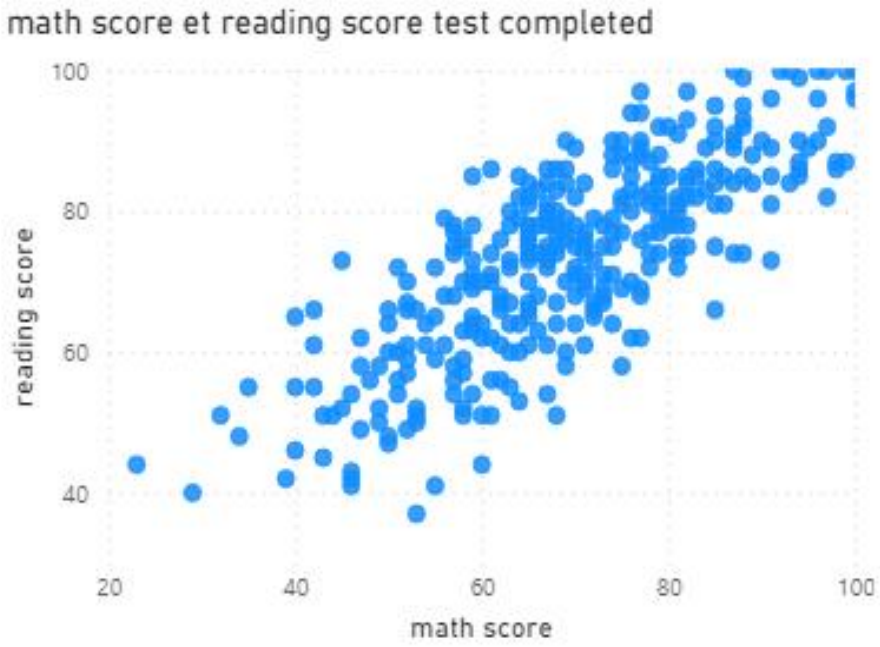


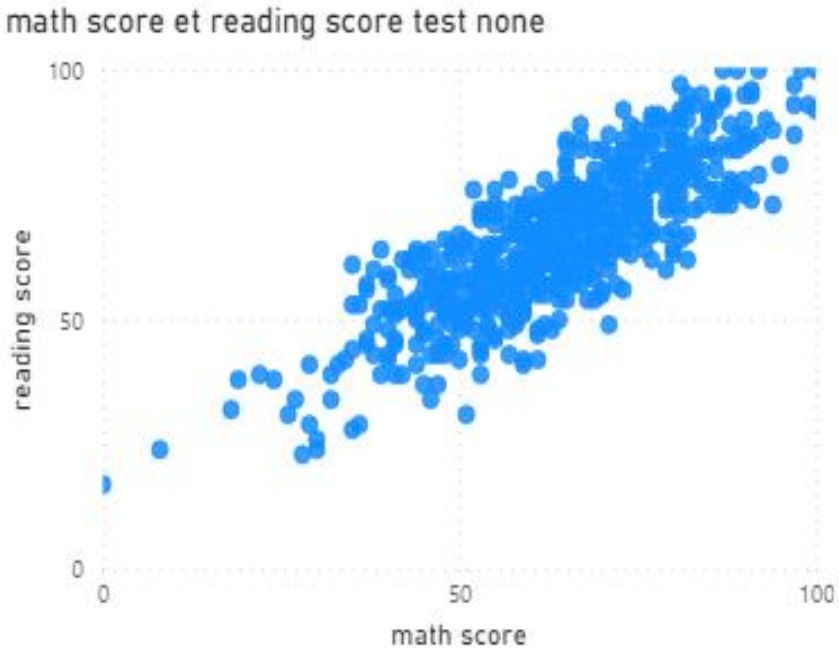


Nous terminons cette partie avec une dernière comparaison à savoir, les étudiants qui ont complété des cours de préparation et ceux qui ne l'ont pas fait.

A notre grande surprise, ceux qui ont complété les cours de préparation n'ont pas été régulier dans les résultats des deux matières, nous disons cela parce que le deuxième nuage de points qui montre ceux qui n'ont complété les cours de préparation a une plus forte corrélation puisque ses points sont moins dispersés que le premier.

Cela rejoint bien notre hypothèse qui n'était pas validée.





Sous Python

Dans cette dernière partie d'analyse, nous avons codé quelques programmes pour faire une analyse en modélisation linéaire simple. Les images qui suivent seront donc des nuages de points avec des droites de régression pour montrer la force et la direction de la relation entre les variables, ces visuels seront accompagnés de leur code Python.

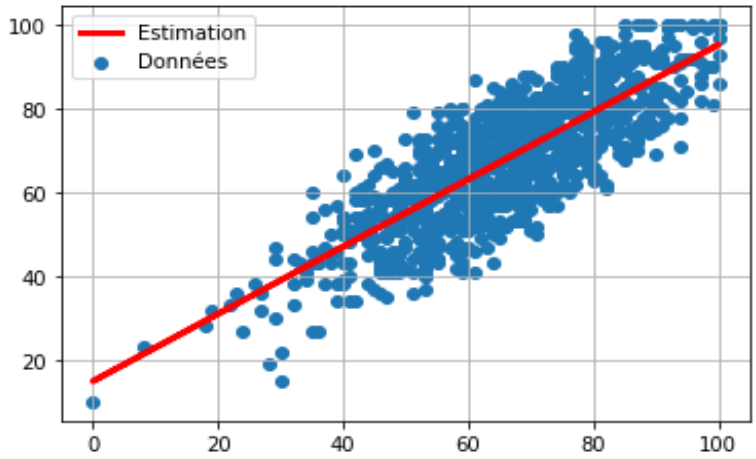
Résultat du premier programme :

R^2 : 0,64

Le résultat de corrélation n'est pas optimal pour estimer la note de maths en fonction de la note d'écriture et inversement

```
analyse(ScoreMaths,ScoreEcriture)
```

```
Moyenne des x = 66.089
Moyenne des y = 68.054
Ecart-type des x = 15.155496659628165
Ecart-type des y = 15.188057281956757
achapeau = 0.8043664714246145
bchapeau = 14.894224270018654
ychapeau = [72.8086102125909, 70.39551079831705, 87.28720669823396, 52.
Résidus =  [1.1913897874090935, 17.604489201682952, 5.712793301766041,
Moyenne résidus =  -5.613287612504792e-16
RCarré = 0.6442342539264908
L'estimation de l'écart-type de l'erreur est : 9.06815680427808
```



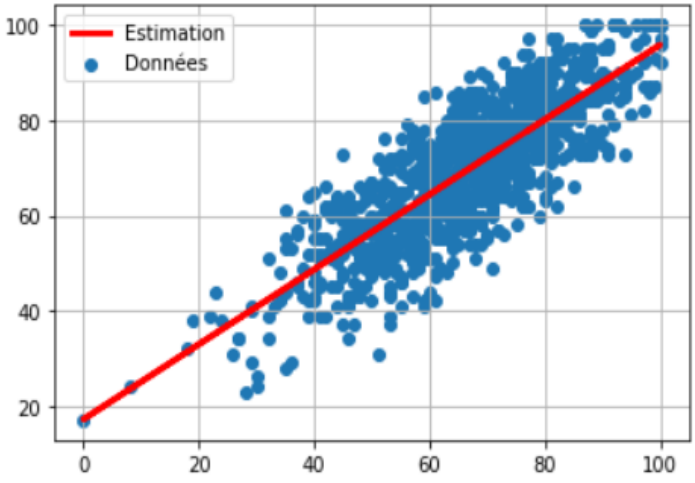
Résultat du deuxième programme :

R^2 : 0,69

Le résultat de corrélation n'est toujours pas encore optimal pour estimer la note de maths en fonction de la note de lecture et inversement

```
[ ]  analyse(ScoreMaths,ScoreLecture)
```

```
Moyenne des x = 66.089
Moyenne des y = 69.169
Ecart-type des x = 15.155496659628165
Ecart-type des y = 14.59289001534652
achapeau = 0.7872292395756425
bchapeau = 17.14180678568536
ychapeau = [73.82231203513163, 71.4606243164047, 87.99243834749319,
Résidus =  [-1.822312035131631, 18.539375683595296, 7.00756165250680
Moyenne résidus =  1.3073986337985843e-15
RCarré = 0.668436506450105
L'estimation de l'écart-type de l'erreur est : 8.411227742671588
```



Résultat du troisième programme :

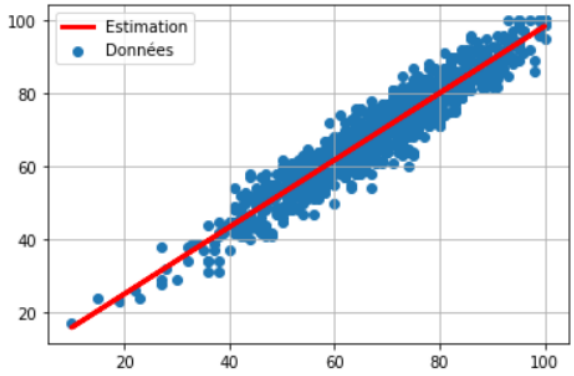
R^2 : 0,91

Estimation écart-type de l'erreur : 4 (soit 2 fois moins que les précédentes analyses)

Le résultat de corrélation est optimal pour estimer la note de l'écriture en fonction de la note de lecture et inversement

```
[ ]  analyse(ScoreEcriture,ScoreLecture)
```

```
Moyenne des x = 68.054
Moyenne des y = 69.169
Ecart-type des x = 15.188057281956757
Ecart-type des y = 14.59289001534652
achapeau = 0.9171906906886343
bchapeau = 6.750504735875673
ychapeau = [74.6226158468346, 87.46328551647551, 92.04923896991866,
Résidus =  [-2.6226158468346057, 2.5367144835244915, 2.9507610300813
Moyenne résidus =  5.89039927945123e-15
RCarré = 0.9112574888913156
L'estimation de l'écart-type de l'erreur est : 4.351529132979537
```



La corrélation de Pearson montre en revanche de bien meilleurs résultats, à la différence du R^2 qui calcule la corrélation entre les notes brutes et les notes estimées.

On observe le même ordre de relations entre R^2 et la corrélation de Pearson

```
print("Corrélation de Pearson entre les scores de Maths et Ecriture = " + str(pearsonr(ScoreMaths,ScoreEcriture)[0]))
print("Corrélation de Pearson entre les scores de Maths et Lecture = " + str(pearsonr(ScoreMaths,ScoreLecture)[0]))
print("Corrélation de Pearson entre les scores de Lecture et Ecriture = " + str(pearsonr(ScoreLecture,ScoreEcriture)[0]))
```

```
Corrélation de Pearson entre les scores de Maths et Ecriture = 0.8026420459498078
Corrélation de Pearson entre les scores de Maths et Lecture = 0.8175796636720539
Corrélation de Pearson entre les scores de Lecture et Ecriture = 0.9545980771462479
```

CONCLUSION

Nous avons validé la plupart de nos hypothèses grâce à différentes techniques d'analyse :

Les mathématiques sont moins bien réussies que les autres matières au profit des tests écrits, les femmes ont bien une régularité meilleure dans les études par rapport aux hommes, elles réussissent donc en général mieux aux examens pourtant, les hommes ont un esprit plus scientifique en effet, ils ont de meilleurs résultats en mathématiques.

Nous avons également confirmé que plus les parents des étudiants ont un niveau d'étude élevé, plus leurs notes sont élevées, par ailleurs, ces étudiants complètent rarement les cours de préparation même si nous avons remarqué que ceux qui le complètent réussissent mieux.

Nous allons clôturer cette étude analytique par les limites que nous avons rencontrées.

Pour commencer, nous n'avons utilisé qu'une seule base de données, elle ne contenait que 1000 lignes pour 8 variables. Nous aurions aimé travailler sur plus de données et de sources différentes.

Ensuite, viens la difficulté de démarrer un travail sur une nouvelle plateforme (Google Colab) tandis que nous maitrisons déjà Jupyter Notebook qui fait la même chose, au final nous avons réussi et cela nous a été plutôt bénéfique.