25/10/2023 | Par : Abderrahmane BOUFARES

Rapport statistique

Analyse de base de données

Introduction

Présentation de la problématique.

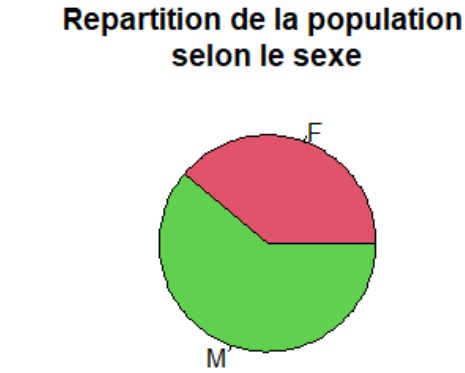
Problématique :

Les rectorats des universités du Sud-Ouest (UPPA, Bordeaux et Toulouse), cherche a établir des règles de décision en modélisation pour chaque licence, la probité de réussite a partir des données observées sur la population étudiante des années précédentes. Afin d'envoyer aux futurs étudiants, des avis reflétant leurs chances de réussite dans la licence choisie, grâce a un programme d'orientation active de lycéens. Notre base de donnée destinée a l'étude se compose de deux type de variables, **des variables** qualitative tel que le sexe (Féminin, Masculin), la profession des parents (Défavorable, Moyen, Favorable et Très Favorable), le retard de l'obtention du BAC (Bac réussi des la première année, réussi en deuxième année ou réussi en troisième année), la série du BAC (ES, L, S, SMS, STI ou STT), type du BAC (General ou Technologique), la qualité de la mention obtenue au BAC (A,B,C,D ou P), l'obtention du BAC (Obtenu ou non obtenu), la mention (Obtenue ou non obtenue), la réussite en première de licence (Réussie ou non réussie) et **une variable quantitative**, celle de la note obtenue a l'épreuve de Math du BAC, entre 1 et 19. Pour faire cette étude, on va travailler avec le logiciel Rstudio et une base de donnée sous R, de nom : L1eco.rdata. Ce logiciel nous permettra de faire des tables de contingences, des tests statistiques sur cette base de données et créer des graphiques, sous forme d'histogramme, diagramme en barre, circulaire et mosaïque, boite a moustache. L'objectif de cette étude, est d'estimer la dépendance et la relation entre les variables croisées, elle permet aussi de comparer l'intensité du lien entre les variables croisées et leurs corrélations, elle nous permet aussi de vérifier l'Independence entre les variables.

Développement

Analyse univariée :

- Sexe :



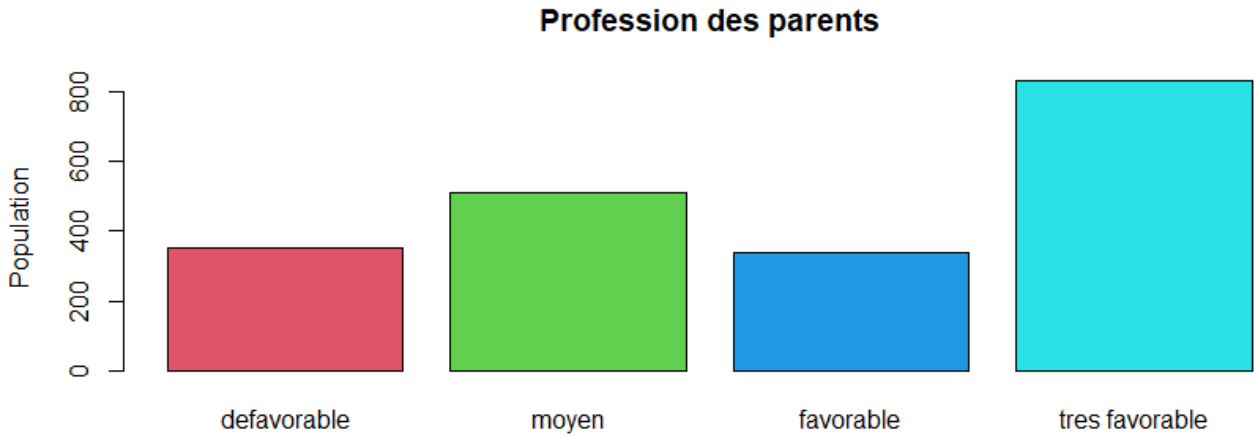
F	M
0.3908329	0.6091671

D'après le diagramme circulaire et la table de contingence on distingue, que notre population est divisée en deux niveaux, le sexe Féminin qui représente 39% de l'échantillon et le sexe Masculin qui représente 60% de l'échantillon.

Développement

Analyse univarié :

- Profession des parents :



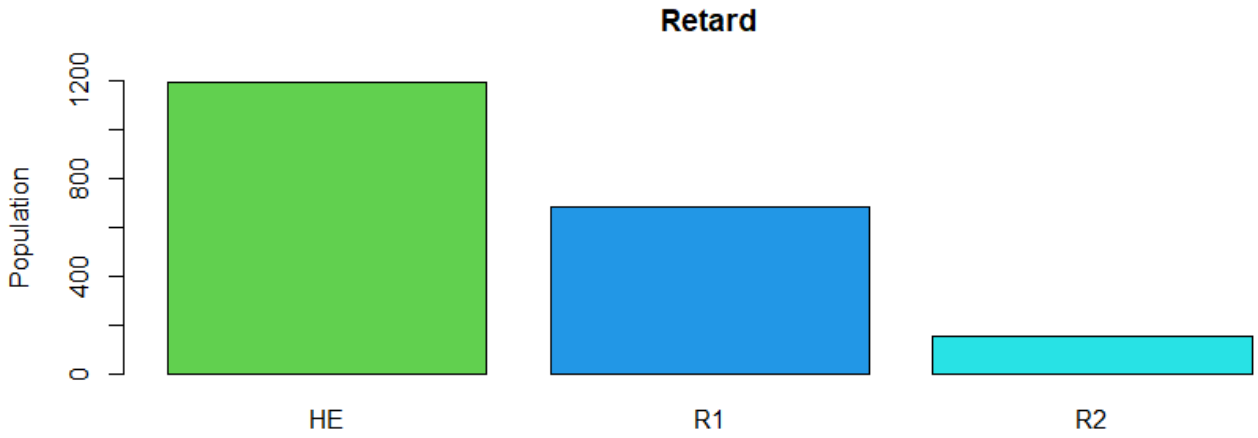

```
pcsf
   defavorable          moyen      favorable tres favorable
     0.1734845      0.2513553      0.1660917      0.4090685
```

D'après le diagramme en barre et la table de contingence on distingue, que notre population est divisée en quatre niveaux, la profession défavorable qui représente 17% de l'échantillon, la profession moyen qui représente 25% de l'échantillon, , la profession favorable qui représente 16% de l'échantillon et la profession très favorable qui représente 40% de l'échantillon

Développement

Analyse univarié :

- Retard :



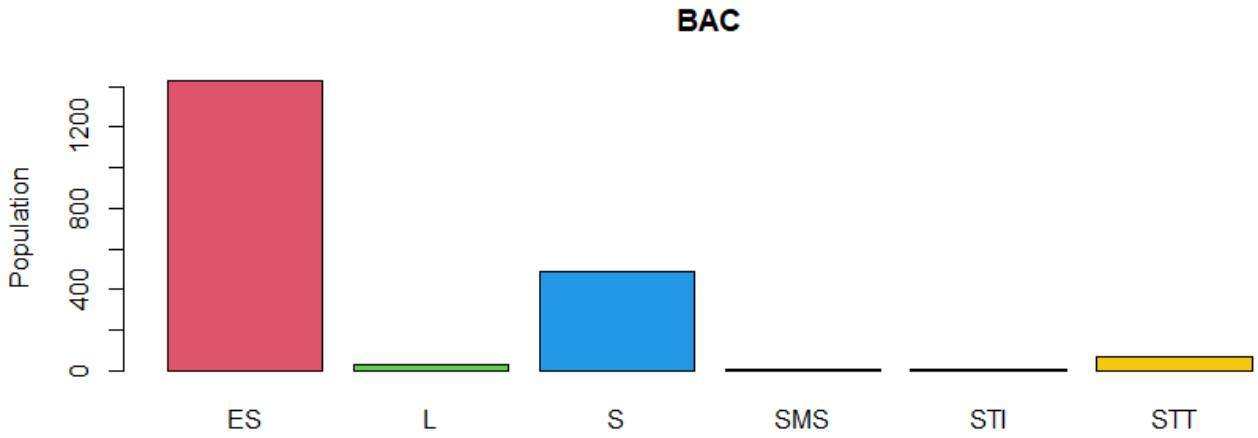
HE	R1	R2
0.58748152	0.33760473	0.07491375

D'après le diagramme en barre et la table de contingence on distingue, que notre population est divisée en trois niveaux, les non redoublants qui représente 58% de l'échantillon, les étudiants qui ont redoubles une seule fois qui représente 33% de l'échantillon, les étudiants qui ont redoubles deux fois qui représente 16% de l'échantillon et la profession très favorable qui représente 40% de l'échantillon.

Développement

Analyse univarié :

- Série du BAC :



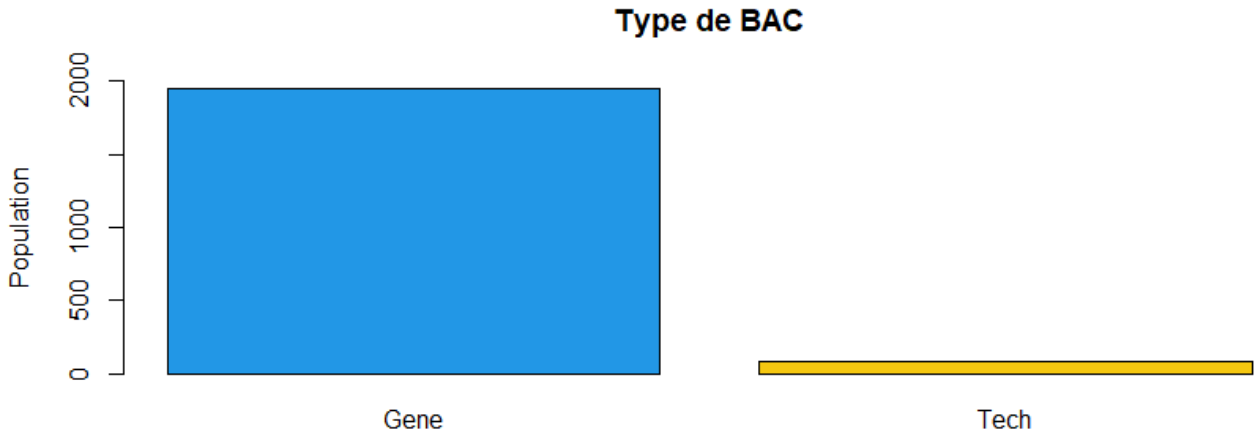
ES	L	S	SMS	STI	STT
0.703302119	0.015278462	0.241005421	0.003449975	0.003942829	0.033021193

D'après le diagramme en barre et la table de contingence on distingue, que notre population est divisée en six niveaux, les séries du BAC : ES, L, S, SMS, STI et STT qui représente respectivement 70,3%, 1,5%, 24,1%, 0,34%, 0,39% et 3,3% de l'échantillon.

Développement

Analyse univarié :

- Type de BAC :



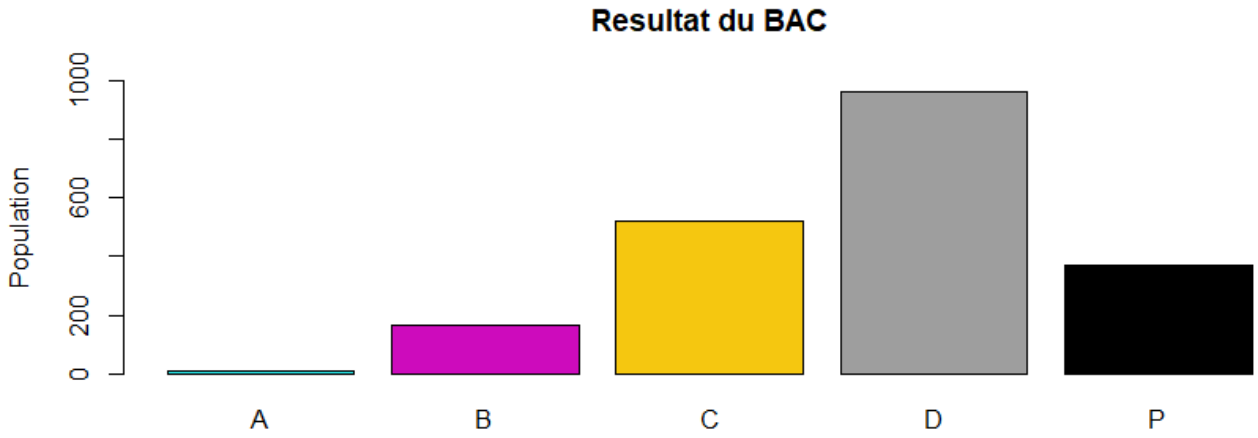

```
    Gene     Tech
0.959586 0.040414
```

D'après le diagramme en barre et la table de contingence on distingue, que notre population est divisée en deux niveaux, les BACS généraux qui représente 96% de l'échantillon, les BACS techniques qui représente 4% de l'échantillon.

Développement

Analyse univarié :

- Qualité de mention :



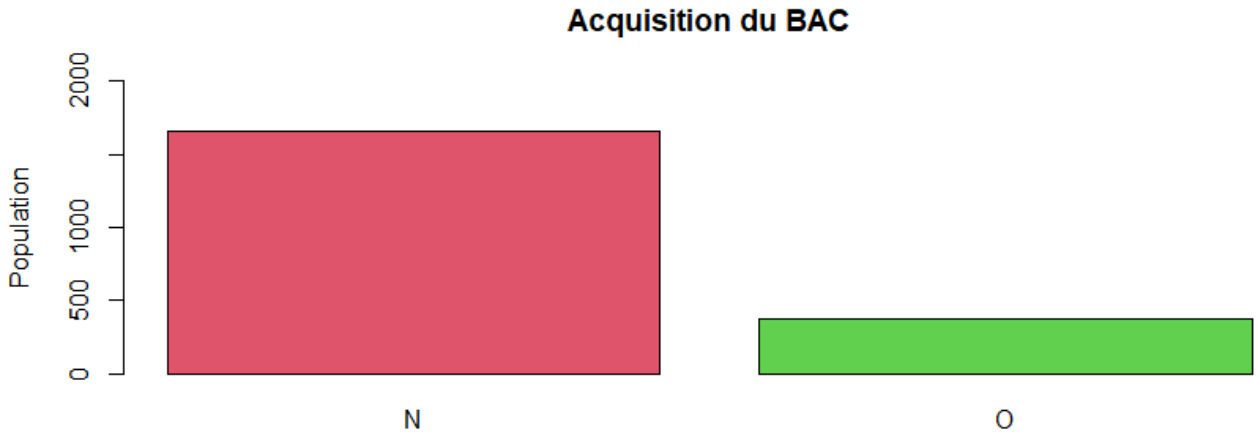
A	B	C	D	P
0.003449975	0.080335140	0.256776737	0.476096599	0.183341548

D'après le diagramme en barre et la table de contingence on distingue, que notre population est divisée en cinq niveaux, les mention BAC : A, B, C, D et P, qui représente respectivement 0,034%, 8%, 25%, 0,47%, 0,39% et 18% de l'échantillon.

Développement

Analyse univarié :

- Obtention du BAC :



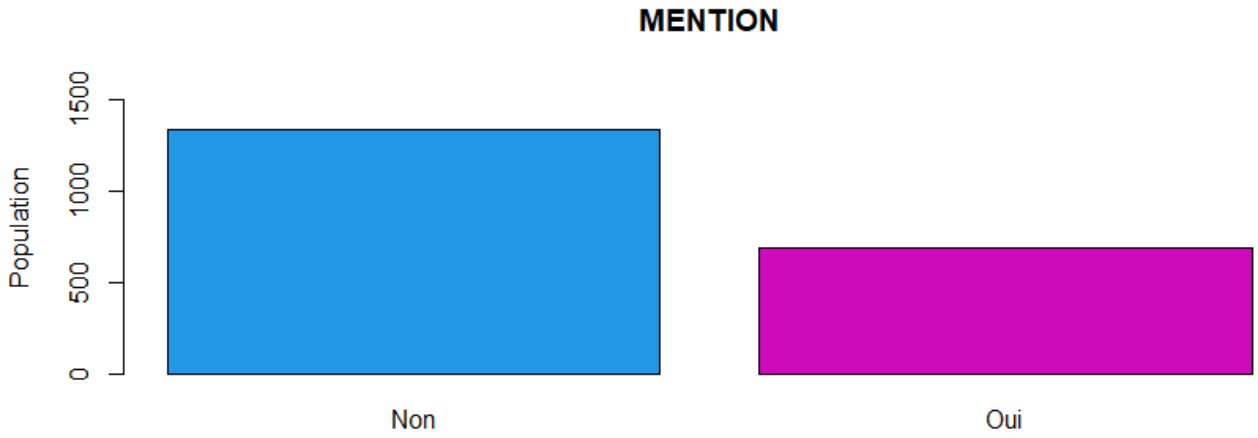
N	O
0.8161656	0.1838344

D'après le diagramme en barre et la table de contingence on distingue, que notre population est divisée en deux niveaux, 81% ont obtenu le BAC et 18% ne l'ont pas obtenu.

Développement

Analyse univarié :

- Obtention de mention :



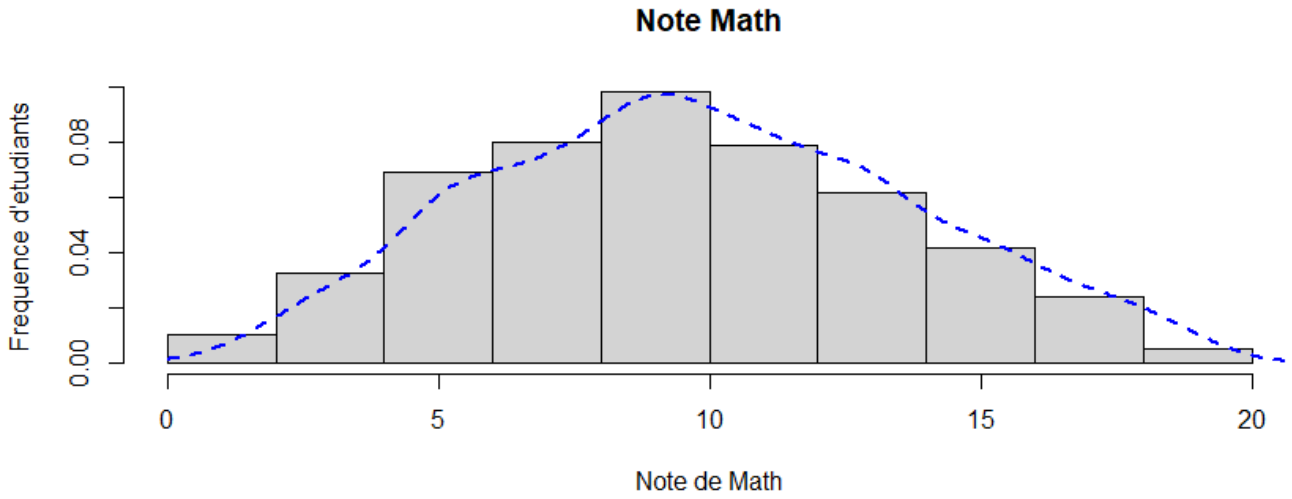

```
      Non       Oui
0.6594381 0.3405619
```

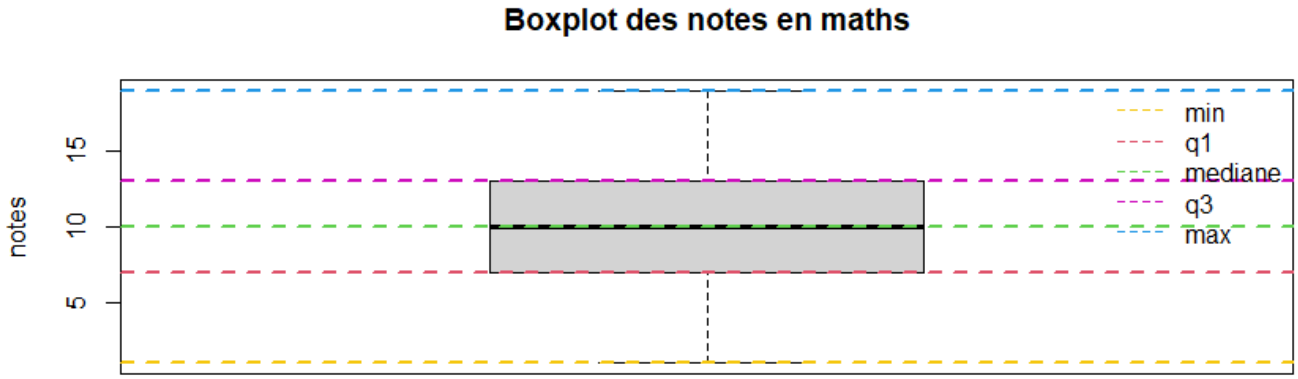
D'après le diagramme en barre et la table de contingence on distingue, que notre population est divisée en deux niveaux, 34% ont obtenu la mention au BAC et 66% ne l'ont pas obtenu.

Développement

Analyse univarié :

- Résultat de l'épreuve de Math du BAC :





	N6	N5	N4	N3	N2	N1
	0.15075377	0.42613065	0.15829146	0.12412060	0.08341709	0.05728643
Intervalle de note	[1,5]	[5,10]	[10,12]	[12,14]	[14,16]	[16,19]

Développement

Analyse univarié :

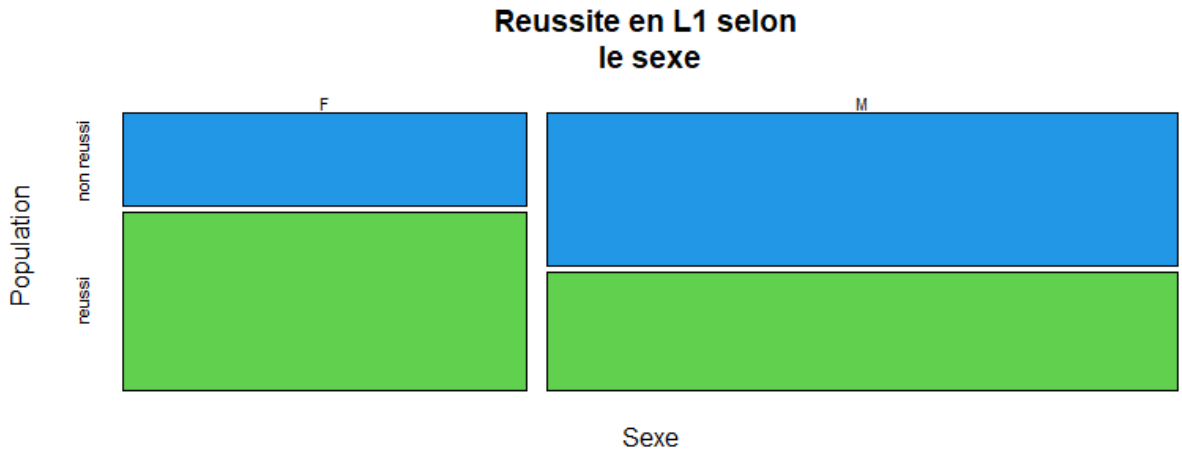
- Résultat de l'épreuve de Math du BAC :

D'après l'histogramme, la boite a moustaches, ainsi que la table de contingence, on déduit que les notes des élèves étaient entre [1,19], avec un minimum de 1/20 et maximum de 19/20. Avec une moyenne de 9,83, médiane de 10, écart type de 3,96 et coefficient de variance égale à 0,40. J'ai découpé cet intervalle, en cinq intervalles, qui corresponde a la qualité de mention. Mon premier intervalle est le [1,5] qui représente 15,07% de l'ensemble des notes, [5,10] qui représente 42,61% de l'échantillon, [10,12] qui représente 15,82%, [12,14] qui représente 12,41%, [14,16] qui représente 8,34% et [16,19] qui représente 5,72%. Ce qui correspond a l'histogramme sous forme de pyramide, vu que l'intervalle le plus proche rassemble plus des deux cinquièmes des notes.

Développement

Analyse entre les valeurs qualitatives :

- Réussite en L1 selon le sexe :



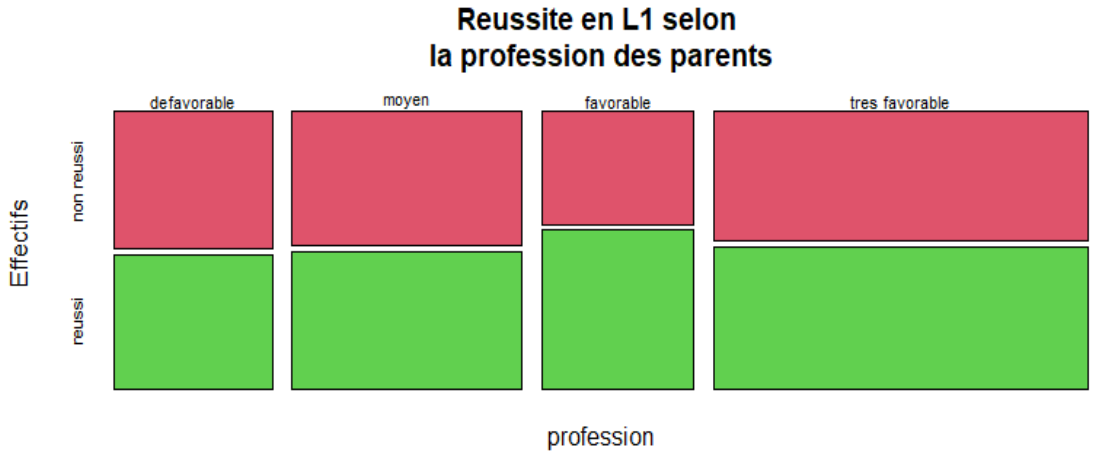
	non reussi	reussi
F	0.3404792	0.6595208
M	0.5622977	0.4377023

Les mosaïques et la table de contingence précèdent nous montre que les femmes ont plus de chance à réussir en L1 que les hommes avec un pourcentage respectivement de 69% et 43%. On a procédé au test KHI2, afin d'être sur qu'il y a une dépendance ou indépendance entre ses deux variables et on a eu un résultat de **p-value inferieure a $2,2^{e}$-16**, ce qui est une valeur très faible par rapport au niveau de signification de 5%, ce qui implique que les deux variables **sont dépendantes**. Pour comprendre l'intensité entre ses deux valeurs, on a utilisé le test du v de Cramer, on a obtenu le résultat suivant : 0,216, ce qui explique qu'il y a vrai une dépendance entre ses variables, mais elle est à **faible intensité** car le résultat du test est inferieur a 50%.

Développement

Analyse entre les valeurs qualitatives :

- Réussite en L1 selon la profession des parents :



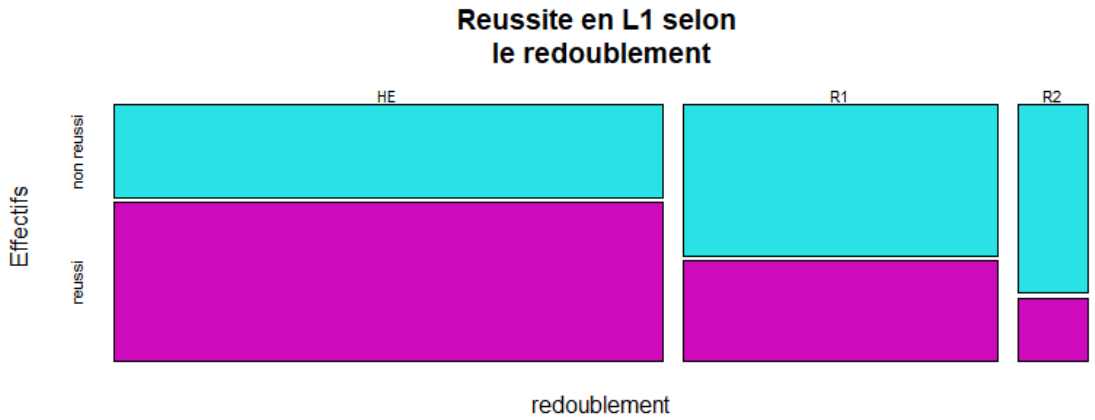
pcsf	reu non reussi	reussi
defavorable	0.5056818	0.4943182
moyen	0.4941176	0.5058824
favorable	0.4154303	0.5845697
tres favorable	0.4759036	0.5240964

Les mosaïques et la table de contingence précèdent, nous montre que la profession des parents n'a pas d'influence sur la réussite en L1. On a procédé au test KHI2, afin d'être sûr qu'il y a une dépendance ou indépendance entre ses deux variables et on a eu un résultat de **p-value égale a 0,07**, ce qui est une valeur supérieure au niveau de signification de 5%, ce qui implique que les deux variables **sont indépendantes**. Pour comprendre l'intensité entre ses deux valeurs, on a utilisé le test du v de Cramer, on a obtenu le résultat suivant : 0,058.

Développement

Analyse entre les valeurs qualitatives :

- Réussite en L1 selon le redoublement au BAC :



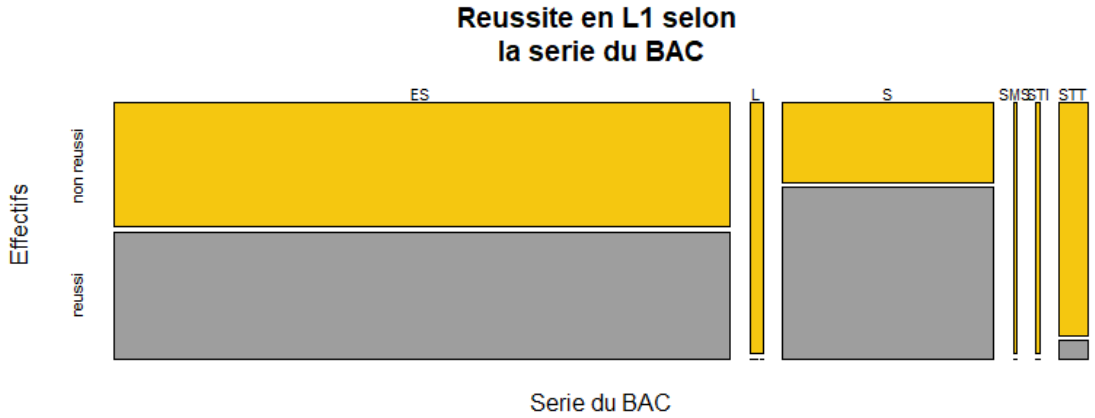

```
  reu
    non reussi    reussi
HE  0.3691275 0.6308725
R1  0.6000000 0.4000000
R2  0.7500000 0.2500000
```

Les mosaïques et la table de contingence précèdent nous montre que les étudiants qui redouble, on moins de chance à réussir par rapport aux étudiants qui ont obtenus leurs BAC dès la première année. On a procédé au test KHI2, afin d'être sûr qu'il y a une dépendance ou indépendance entre ses deux variables et on a eu un résultat de **p-value inferieure a $2,2^{e}$-16**, ce qui est une valeur très faible par rapport au niveau de signification de 5%, ce qui implique que les deux variables **sont dépendantes**. Pour comprendre l'intensité entre ses deux valeurs, on a utilisé le test du v de Cramer, on a obtenu le résultat suivant : 0,265, ce qui explique qu'il y a une dépendance entre ses variables, mais **son intensité est faible**, car le résultat du test est inferieur a 50%.

Développement

Analyse entre les valeurs qualitatives :

- Réussite en L1 selon la série du BAC :



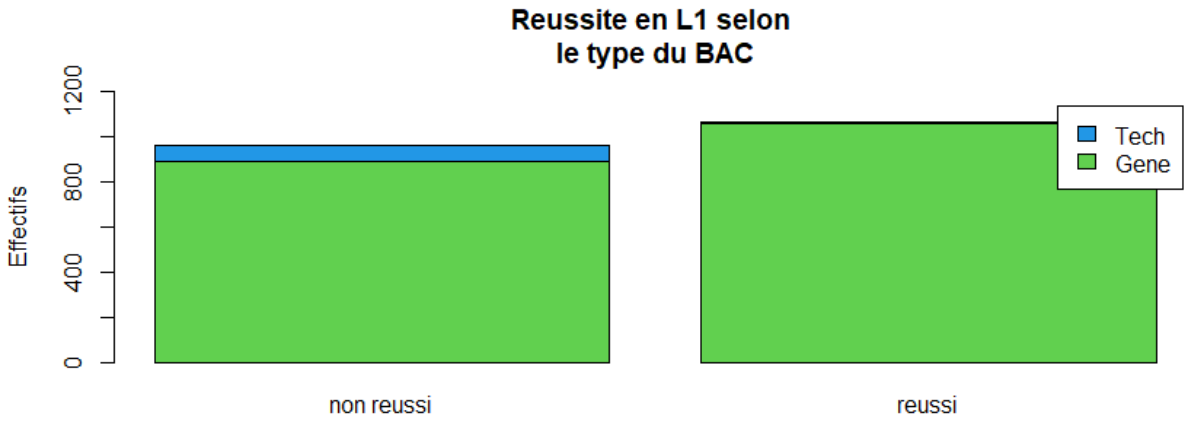
	reu	
	non reussi	reussi
ES	0.49264191	0.50735809
L	1.00000000	0.00000000
S	0.31492843	0.68507157
SMS	1.00000000	0.00000000
STI	1.00000000	0.00000000
STT	0.92537313	0.07462687

Les mosaïques et la table de contingence précèdent nous montre que les étudiants de la série du BAC (S et STT), on distingue aussi que ceux de la série (L, SMS et STI) n'ont aucune chance de réussir. On a procédé au test KHI2, afin d'être sûr qu'il y a une dépendance ou indépendance entre ses deux variables et on a eu un résultat de **p-value inferieure a $2,2^{e}$-16**, ce qui est une valeur très faible par rapport au niveau de signification de 5%, ce qui implique que les deux variables **sont dépendantes**. Pour comprendre l'intensité entre ses deux valeurs, on a utilisé le test du v de Cramer, on a obtenu le résultat suivant : 0,28, ce qui explique qu'il y a vrai une dépendance entre ses variables, mais elle est à **faible intensité** car le résultat du test est inferieur a 50%.

Développement

Analyse entre les valeurs qualitatives :

- Réussite en L1 selon le type du BAC :



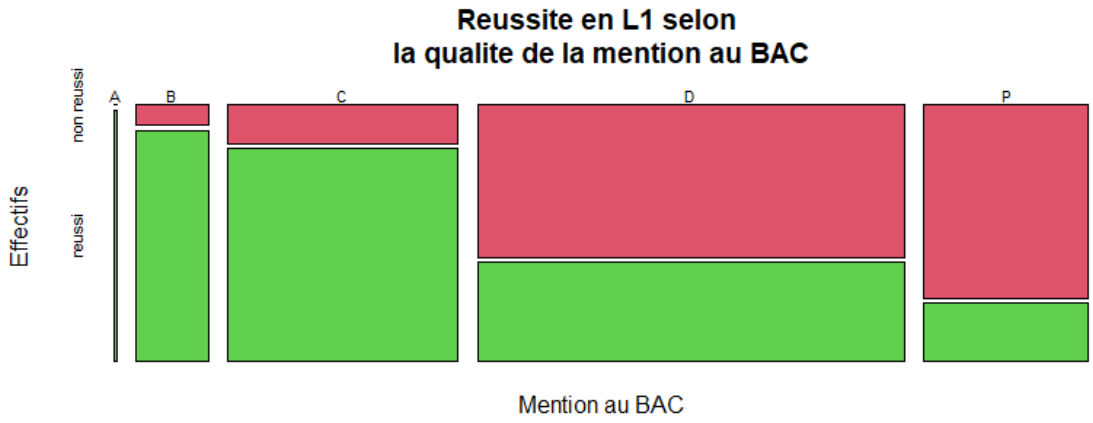

```
reu
     non reussi     reussi
Gene 0.45608629 0.54391371
Tech 0.93902439 0.06097561
```

Le digramme en barre et la table de contingence précèdent nous montre que les étudiants du type BAC général ont plus de chance à réussir en L1 que les autres avec un pourcentage respectivement de 54% et 6%. On a procédé au test KHI2, afin d'être sûr qu'il y a une dépendance ou indépendance entre ses deux variables et on a eu un résultat de **p-value inferieure a 2,2^{e}-16**, ce qui est une valeur très faible par rapport au niveau de signification de 5%, ce qui implique que les deux variables **sont dépendantes**. Pour comprendre l'intensité entre ses deux valeurs, on a utilisé le test du v de Cramer, on a obtenu le résultat suivant : 0,19, ce qui explique qu'il y a une dépendance entre ses variables, mais elle est à **faible intensité**, car le résultat du test est inferieur a 50%.

Développement

Analyse entre les valeurs qualitatives :

- Réussite en L1 selon la qualité de la mention du BAC :



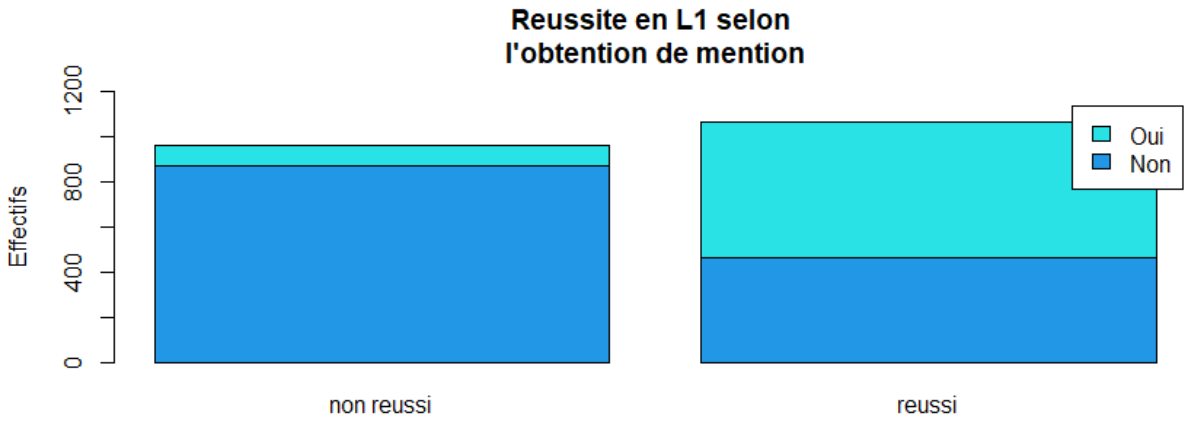

```
reu
  non reussi    reussi
A  0.0000000 1.0000000
B  0.0797546 0.9202454
C  0.1516315 0.8483685
D  0.6076605 0.3923395
P  0.7688172 0.2311828
```

Les mosaïques et la table de contingence précèdent nous montre que plus la qualité de mention au BAC est meilleure, plus l'étudiant a de chance pour réussir son L1. On a procédé au test KHI2, afin d'être sûr qu'il y a une dépendance ou indépendance entre ses deux variables et on a eu un résultat de **p-value inferieure a $2,2^{e}$-16**, ce qui est une valeur très faible par rapport au niveau de signification de 5%, ce qui implique que les deux variables **sont dépendantes**. Pour comprendre l'intensité entre ses deux valeurs, on a utilisé le test du v de Cramer, on a obtenu le résultat suivant : 0,5, ce qui explique qu'il y a vrai une dépendance entre ses variables, avec une **moyen intensité** entre eux, car le résultat du test est égal a 50%.

Développement

Analyse entre les valeurs qualitatives :

- Réussite en L1 selon l'obtention de la mention au BAC :



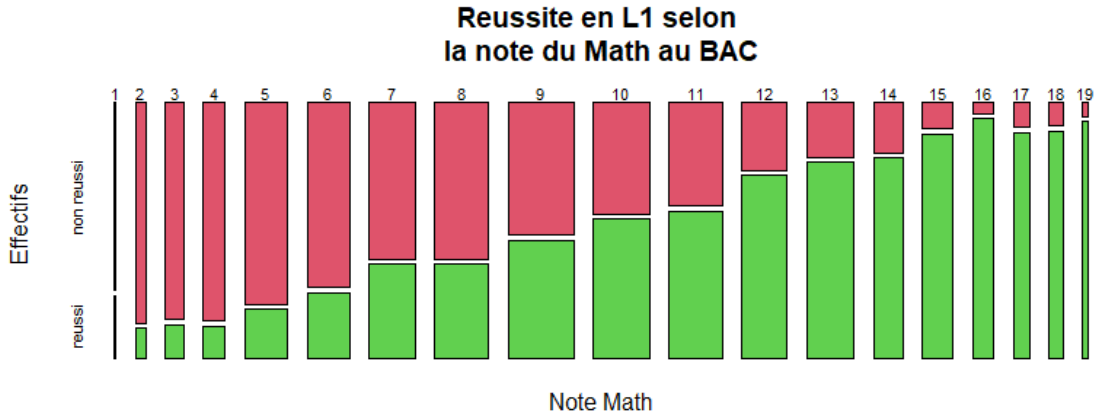

```
reu
     non reussi    reussi
Non   0.6524664 0.3475336
Oui   0.1331404 0.8668596
```

Le diagramme en barre et la table de contingence précèdent nous montre que les étudiants qui ont obtenu leur BAC avec mention sont les plus favorisé à réussir leur L1. On a procédé au test KHI2, afin d'être sûr qu'il y a une dépendance ou indépendance entre ses deux variables et on a eu un résultat de **p-value inferieure a $2{,}2^{e}$-16**, ce qui est une valeur très faible par rapport au niveau de signification de 5%, ce qui implique que les deux variables **sont dépendantes**. Pour comprendre l'intensité entre ses deux valeurs, on a utilisé le test du v de Cramer, on a obtenu le résultat suivant : 0,49, ce qui explique qu'il y a vrai une dépendance entre ses variables, avec une **moyen intensité** entre eux, car le résultat du test est égal à 50%.

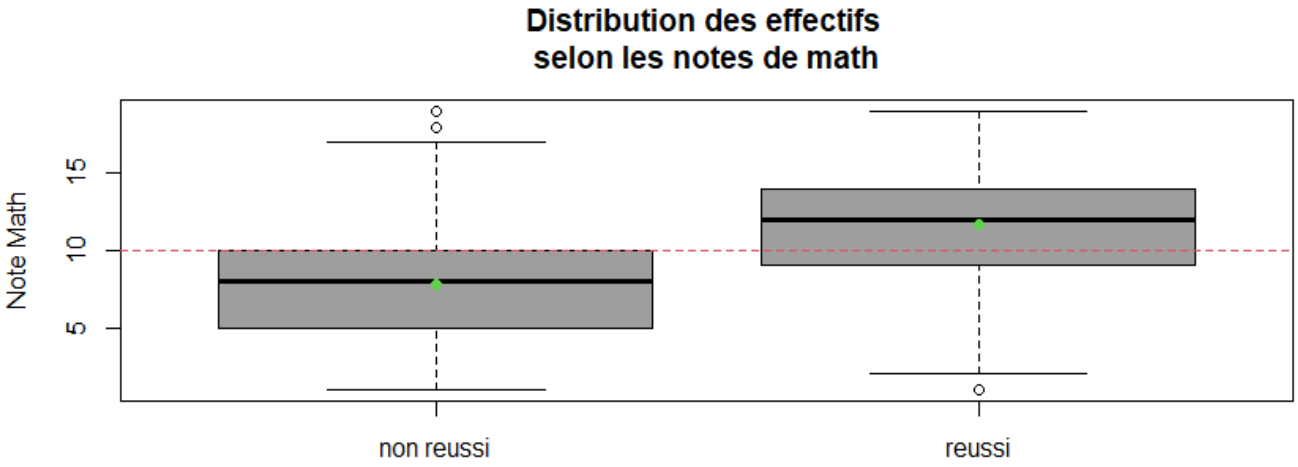
Développement

Analyse entre une variable qualitative et une variable quantitative :

- Réussite en L1 selon l'obtention de la note de l'épreuve de Math au BAC :



	non reussi	reussi
1	0.75000000	0.25000000
2	0.87878788	0.12121212
3	0.86440678	0.13559322
4	0.87142857	0.12857143
5	0.80434783	0.19565217
6	0.73722628	0.26277372
7	0.62328767	0.37671233
8	0.62427746	0.37572254
9	0.52857143	0.47142857
10	0.44505495	0.55494505
11	0.41176471	0.58823529
12	0.26896552	0.73103448
13	0.21710526	0.78289474
14	0.20000000	0.80000000
15	0.10309278	0.89690722
16	0.04347826	0.95652174
17	0.09615385	0.90384615
18	0.09302326	0.90697674
19	0.05263158	0.94736842



Développement

- Analyse entre une variable qualitative et une variable quantitative :
- Réussite en L1 selon l'obtention de la note de l'épreuve de Math au BAC :

Les mosaïques, la boite a moustache et la table de contingence précèdent nous montre que les étudiants ayant les bonnes notes à l'épreuve des maths au BAC, sont les étudiants qui réussissent mieux en L1, selon la table de contingence précédente, des que la note de l'épreuve augmente, la fréquence liée a la réussite augmente et vice versa. Via la boite a moustache, on observe que les trois quarts des étudiants qui ont eu moins ou égale à 10 à l'épreuve des math au BAC, n'ont pas pu réussir leur première année. On a procédé au test KHI2, afin d'être sûr qu'il y a une dépendance ou indépendance entre ses deux variables et on a eu un résultat de **p-value inferieure a $2,2^{e}$-16**, ce qui est une valeur très faible par rapport au niveau de signification de 5%, ce qui implique que les deux variables **sont dépendantes**. Pour comprendre l'intensité entre ses deux valeurs, on a utilisé le test du v de Cramer, on a obtenu le résultat suivant : 0,5, ce qui explique qu'il y a vrai une dépendance entre ses variables, avec une **moyen intensité** entre eux, car le résultat du test est égal à 50%.

Conclusion

- Conclusion :

D’après les études sur la base de données, on distingue que la réussite des étudiants en L1 est dépendante de toutes les variables sauf la profession que son influence est presque nul. Parmi ses variables dépendantes y ont à qui sont plus intense que les autres, comme je le montre avec le tableau suivant :

Variable/Réussite	P-value	Etat de la relation	V-cramer	Intensité
Type de BAC	$<2,2^{e}$-16	Dépendance	0,19	Faible
Sexe	$<2,2^{e}$-16	Dépendance	0,216	Faible
Redoublement	$<2,2^{e}$-16	Dépendance	0,265	Faible
Série du BAC	$<2,2^{e}$-16	Dépendance	0.28	Faible
Obtention de mention	$<2,2^{e}$-16	Dépendance	0,49	Moyenne
Qualité de mention	$<2,2^{e}$-16	Dépendance	0,5	Moyenne
Note a l’épreuve de Math au BAC	$<2,2^{e}$-16	Dépendance	0,5	Moyenne
Profession des parents	0,07	Indépendance	0,05	Faible