Chapitre 1 : Prise en main et algèbre

Exercice 13

Un échantillon de dossiers d'enfants a été saisi. Ce sont des enfants vus lors d'une visite en 1ère section de maternelle en 1996-1997 dans des écoles de Bordeaux (Gironde, France). L'échantillon est constitué de 152 enfants âgés de 3 ou 4 ans. Nous Considérons le jeu de donnée Poids-Naissance. Il s'agit ici d'expliquer la variabilité du poids de naissance de l'enfant en fonction des caractéristiques de la mère, de ses antécédents et de son comportement pendant la grossesse. La variable à expliquer est le poids de naissance de l'enfant (variable quantitative BWT, exprimée en grammes) et les facteurs étudiés (variables explicatives) sont : Age de la mère, Poids de la mère lors du dernier cycle menstruel, "Race" de la mère, Tabagisme durant la grossesse, Nombre d'antécédents de prematurité, Antécédents d'hypertension, Présence d'irritabilité utérine, Nombre de visites à un médecin durant le premier trimestre de la grossesse, Poids de naissance et Poids de naissance inférieur ou égal à 2500 g.

Table 1: Extrait des données Poids Naissance

ID	AGE	LWT	RACE	SMOKE	PTL	HT	UI	FVT	BWT	LOW
85	19	182	2	0	0	0	1	0	2523	0
86	33	155	3	0	0	0	0	3	2551	0
87	20	105	1	1	0	0	0	1	2557	0
88	21	108	1	1	0	0	1	2	2594	0
89	18	107	1	1	0	0	1	0	2600	0
91	21	124	3	0	0	0	0	0	2622	0

Le tableau 1 est un court extrait du jeu de données. Ici la variable LWT, qui correspond au poids de la mère, est exprimé en livres, nous la modifions donc pour l'avoir en kilogrammes.

Table 2: Extrait des données avec chagement d'unité du poids de la mère

ID	AGE	LWT	RACE	SMOKE	PTL	HT	UI	FVT	BWT	LOW
85	19	82.554	2	0	0	0	1	0	2523	0
86	33	70.307	3	0	0	0	0	3	2551	0
87	20	47.627	1	1	0	0	0	1	2557	0
88	21	48.988	1	1	0	0	1	2	2594	0
89	18	48.534	1	1	0	0	1	0	2600	0
91	21	56.245	3	0	0	0	0	0	2622	0

Nous obtenons le jeu de données du tableau 2.

Réalisons maintenant quelque tri à plats avec ces données.

Table 3: Tri à plat de l'age de la mère

classes	effectif
(14,20.2]	69
(20.2,26.4]	74
(26.4,32.6]	37
(32.6,38.8]	8

classes	effectif
(38.8,45]	1

Avec l'age de la mère, On remarque dans le tableau 3 que 74 des enfants de nos données ont une mère agé entre 20 et 26 ans.

Table 4: Tri à plat du poids de la mère

classes	effectif
(36.2,51.7]	60
(51.7,67.1]	87
(67.1,82.6]	26
(82.6,98]	12
(98,113]	4

Le tableau 4 nous apprend que 88 mère des enfants de nos données, ont un poid qui se situe entre 51.7 et 67.1 kg.

Table 5: Tri à plat de la race de l'enfant

	Effectif
Blanche	96
Noir	26
Autre	67

Le tableau 5 nous indique que 96 enfants des données on une race dite "Blanche".

Table 6: Tri à plat du tabagisme durant la grossesse

	effectif
Non	115
Oui	74

Et pour finir 115 enfants de nos données, avais une mère qui fumé durant sa grossessesse, comme nous indique le tableau 6.

Exercice 14

Nous allons crée un jeu de donnée personnel, l'objectif sera de voir les manipulation possible sur des données.

Table 7: Jeu de données acteur

Mort.a	Années.de.carrière	Nombre.de.films	Prénom	Nom	Date.du.deces
93	66	211	Michel	Galabru	04-01-2016
53	25	58	André	Raimbourg	23-09-1970
72	48	98	Jean	Gabin	15-10-1976
68	37	140	Louis	De Funès	27-01-1983

Mort.a	Années.de.carrière	Nombre.de.films	Prénom	Nom	Date.du.deces
68	31	74	Lino	Ventura	22-10-1987
53	32	81	Jacques	Villeret	28-01-2005

Le tableau 7 nous illustre le jeu de données. Chaque individu correspond à un acteur ou on retrouve son nom, prenom, son nombre d'année de carriere, son nombre de de film, et la date de sa mort avec l'age.

Table 8: Prénom du jeu de données acteur

Prénom
Michel
André
Jean
Louis
Lino
Jacques

Nous pouvons aussi extraire une colonne en particulier, dans le tableau 8 c'est « prénom » qui est extrait.

Table 9: Donnée acteur trié par l'age du décès

	Age.du.décès	Années.de.carrière	Nombre.de.films	Prénom	Nom	Date.du.deces
2	53	25	58	André	Raimbourg	23-09-1970
6	53	32	81	Jacques	Villeret	28-01-2005
4	68	37	140	Louis	De Funès	27-01-1983
5	68	31	74	Lino	Ventura	22-10-1987
3	72	48	98	Jean	Gabin	15-10-1976
1	93	66	211	Michel	Galabru	04-01-2016

Nous modifions ensuite le nom de la colonne « Mort.à » , en « Age.du.décès », comme le montre le tableau 9. Et pour finir, nous pouvons ordonner le jeu de donnée selon une conditions. Ici nous voulons ordonner par « Age.du.décès » croissant, ce qui est fait dans le tableau 9 .

Exercice 15

Le goût d'un fromage dépend de la concentration de plusieurs composés chimiques, dont : la concentration d'acide acétique (variable X1), la concentration d'hydrogène sulfuré (variable X2), la concentration d'acide lactique (variable X3).

Pour 30 types de fromage, on dispose du score moyen attribué par des goûteurs (caractère Y).

Table 10: Extrait du jeu de donnée fromage

Y	X1	X2	X3
12.3	4.543	3.135	0.86
20.9	5.159	5.043	1.53
39.0	5.366	5.438	1.57
47.9	5.759	7.496	1.81
5.6	4.663	3.807	0.99

Y	X1	X2	X3
25.9	5.697	7.601	1.09

Voici un extrait des données représenté dans le tableau 10 . On retrouve bien un total de 30 individus qui corresponde à des types de fromage. Il bien les 4 variables toute quantitatives. Il y a X1, X2, X3, et Y.

Table 11: Statistiques élémentaires des données fromage

Y	X1	X2	X3
Min. : 0.70	Min. :4.477	Min. : 2.996	Min. :0.860
1st Qu.:13.55	1st Qu.:5.237	1st Qu.: 3.978	1st Qu.:1.250
Median :20.95	Median :5.425	Median : 5.329	Median :1.450
Mean :24.53	Mean :5.498	Mean : 5.942	Mean :1.442
3rd Qu.:36.70	3rd Qu.:5.883	3rd Qu.: 7.575	3rd Qu.:1.667
Max. :57.20	Max. :6.458	Max. :10.199	Max. :2.010

tableau 11 nous montre les statistiques élémentaire pour chacune des variables. Par exemple pour Y on trouve une valeur moyenne de 24,53, un minimum de 0.7 et un maximum de 57.20. Matrice des nuages de points

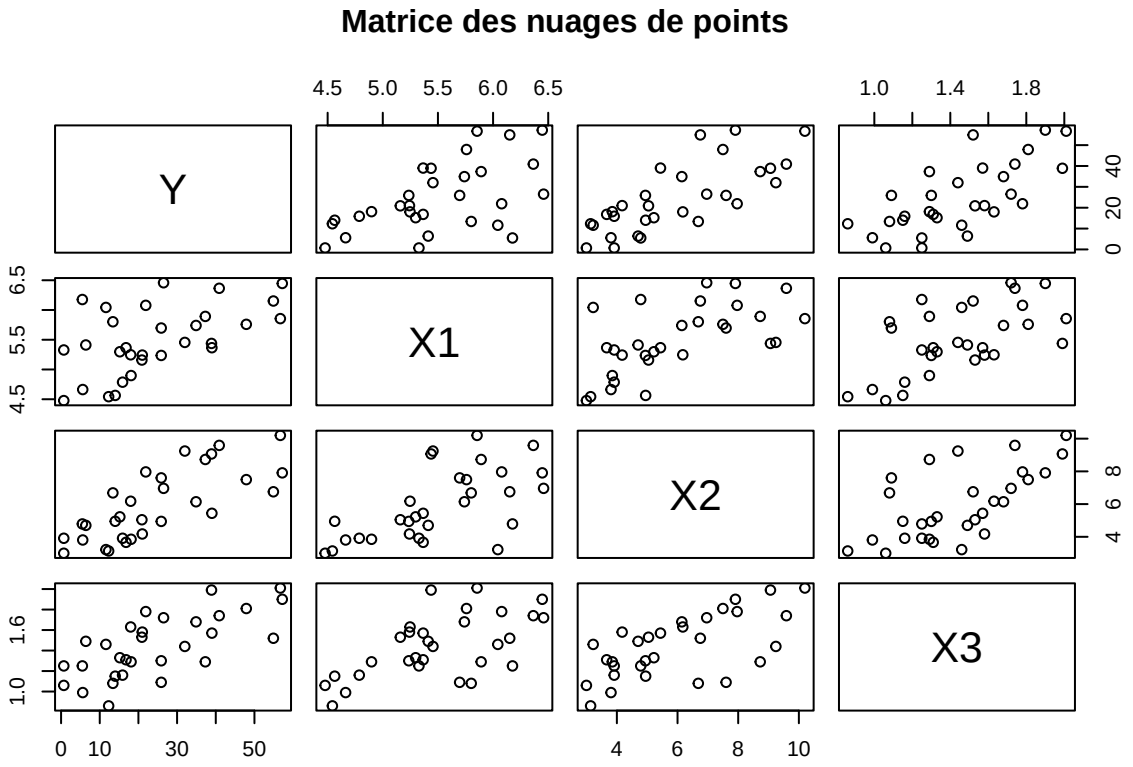

Figure 1: Ozone en fonction des saisons

La figure 1 représente la Matrice de nuage de point entre chacune des variables. Ce sont les nuages de points des croisements deux à deux entre chaque variables de nos données.

Table 12: Extrait des données fromage filtré

	Y	X1	X2	X3
2	20.9	5.159	5.043	1.53
3	39.0	5.366	5.438	1.57
6	25.9	5.697	7.601	1.09
7	37.3	5.892	8.726	1.29
10	21.0	5.242	4.174	1.58
11	34.9	5.740	6.142	1.68

Nous allons maintenant crée avec sous jeu de données avec les contraintes suivante X1 $>$ 5.1 et X3 $<$ 1.77. C'est ce qui est représenté dans le tableau 12 .

Table 13: Statistiques élémentaires des données fromage filtrées

Y	X1	X2	X3
Min. : 0.70	Min. :5.159	Min. :3.219	Min. :1.080
1st Qu.:14.30	1st Qu.:5.313	1st Qu.:4.744	1st Qu.:1.295
Median :21.00	Median :5.455	Median :5.438	Median :1.460
Mean :23.52	Mean :5.654	Mean :5.946	Mean :1.435
3rd Qu.:33.45	3rd Qu.:5.968	3rd Qu.:6.857	3rd Qu.:1.575
Max. :54.90	Max. :6.458	Max. :9.588	Max. :1.740

Apres ce changement on trouve certaine valeur différentes des statistiques élémentaire, par exemple la moyenne de Y est maintenant de 23,52. Nous voyons les nouvelles statistiques dans le tableau 13.

Exercice 16

Les données que utiliserons sont directement implanté dan R, il s'agit des données « airquality ».

Table 14: Extrait des données airquality

Ozone	Solar.R	Wind	Temp	Month	Day
41	190	7.4	67	5	1
36	118	8.0	72	5	2
12	149	12.6	74	5	3
18	313	11.5	62	5	4
NA	NA	14.3	56	5	5
28	NA	14.9	66	5	6

Le tableau 14 nous montre un extrait des données. Il s'agit des relevés quotidiens des valeurs de qualité de l'air suivantes du 1er mai 1973 au 30 septembre 1973. Il y a 153 individus pour 6 variables.

- Ozone : Ozone moyen en parties par milliard de 1300 à 1500 heures à Roosevelt Island.
- R. solaire : rayonnement solaire à Langley dans la bande de fréquence 4000-7700 Angstroms de 0800 à 1200 heures à Central Park.
- Wind : Vitesse moyenne du vent en miles par heure à 0700 et 1000 heures à l'aéroport de LaGuardia.
- Temp : Température maximale quotidienne en degrés Fahrenheit à l'aéroport de La Guardia.

- Day : Le jour

Les données ont été obtenues auprès du New York State Department of Conservation (données sur l'ozone) et du National Weather Service (données météorologiques).

Table 15: Statistique élémentaire des données airquality

Ozone	Solar.R	Wind	Temp
Min. : 1.00	Min. : 7.0	Min. : 1.700	Min. :56.00
1st Qu.: 18.00	1st Qu.:115.8	1st Qu.: 7.400	1st Qu.:72.00
Median : 31.50	Median :205.0	Median : 9.700	Median :79.00
Mean : 42.13	Mean :185.9	Mean : 9.958	Mean :77.88
3rd Qu.: 63.25	3rd Qu.:258.8	3rd Qu.:11.500	3rd Qu.:85.00
Max. :168.00	Max. :334.0	Max. :20.700	Max. :97.00
NA's :37	NA's :7	NA	NA

Le tableau 15 nous montre les statistiques élémentaires sur nos variables quantitatives, ainsi que les valeur manquante. Pour la variable ozone on remarque 37 valeurs manquante et une moyenne de 42.13.

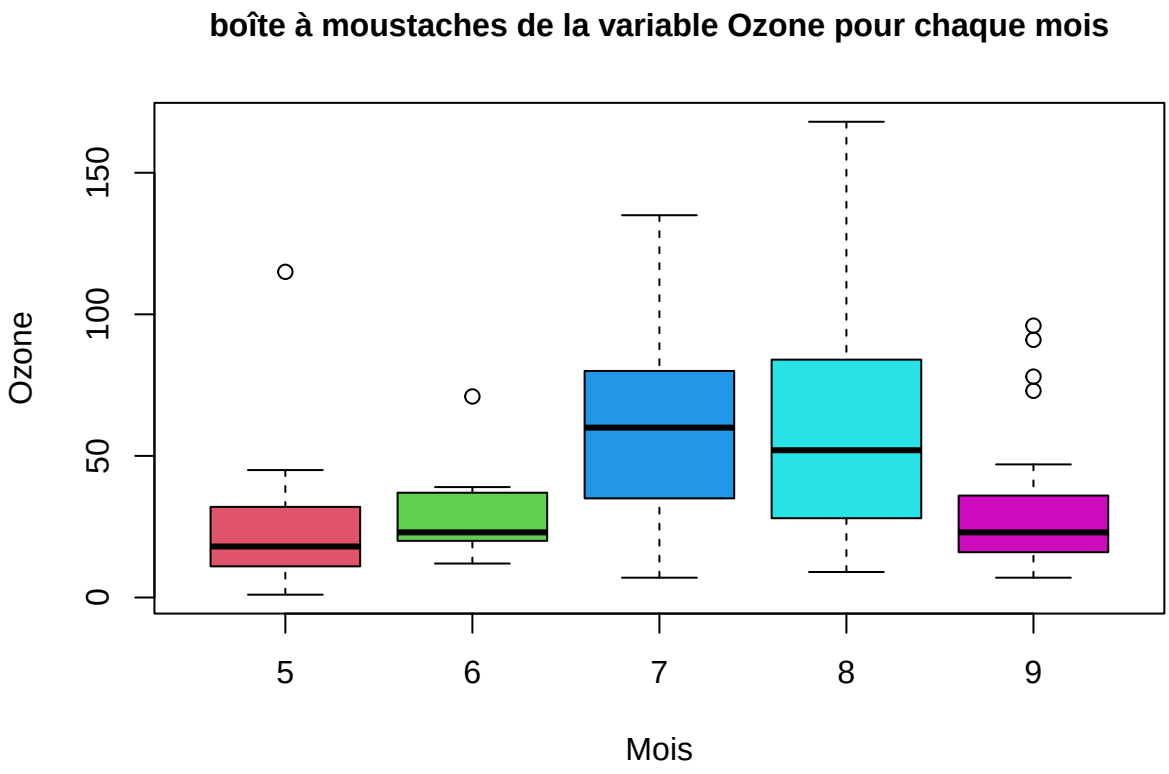

Figure 2: boîte à moustaches de l'Ozone pour chaque mois

On remarque, grâce à la figure 2, des diagrammes à moustache avec une tendance similaire pour les mois 5, 6, et 9, avec des valeur de l'ozone peu élevé qui varie mois. Alors que pour les mois 7 et 8, la valeur de l'ozone sont plus forte et beaucoup plus répartie.

Table 16: Extrait des données airquality avec la saison

Ozone	Solar.R	Wind	Temp	Month	Day	saison
41	190	7.4	67	5	1	printemps
36	118	8.0	72	5	2	printemps
12	149	12.6	74	5	3	printemps
18	313	11.5	62	5	4	printemps
NA	NA	14.3	56	5	5	printemps
28	NA	14.9	66	5	6	printemps

Pour notre analyse, nous rajoutons une variable saison. Le tableau 16 montre nos nouvelle données.

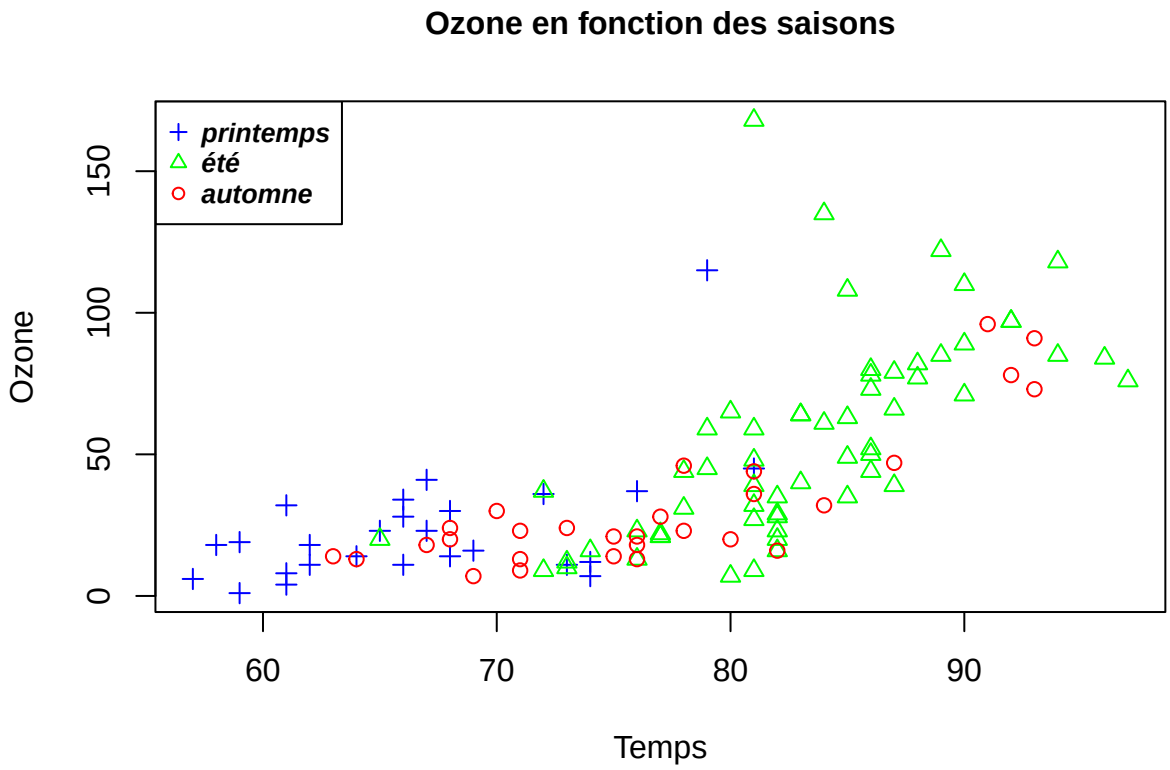

Figure 3: Ozone en fonction des saisons

Avec la figure 3 on remarque qu'il y a une relation positive linéaire croissante entre chaque l'ozone et le temps. Cette relation est présente pour chacune des saisons. Le temps est plus élevé en été et plus faible en hiver. Et comme vu précédemment avec la figure 2, la concentration d'Ozone est plus forte en été, qui correspond aux mois 7 et 8.

Exercice 17

Nous nous intéressons à la fonctions suivant yi = 1.7+2.1 i + ei , i entre 1 et 100 et les ei suivant une lois $N(0,5^2)$.

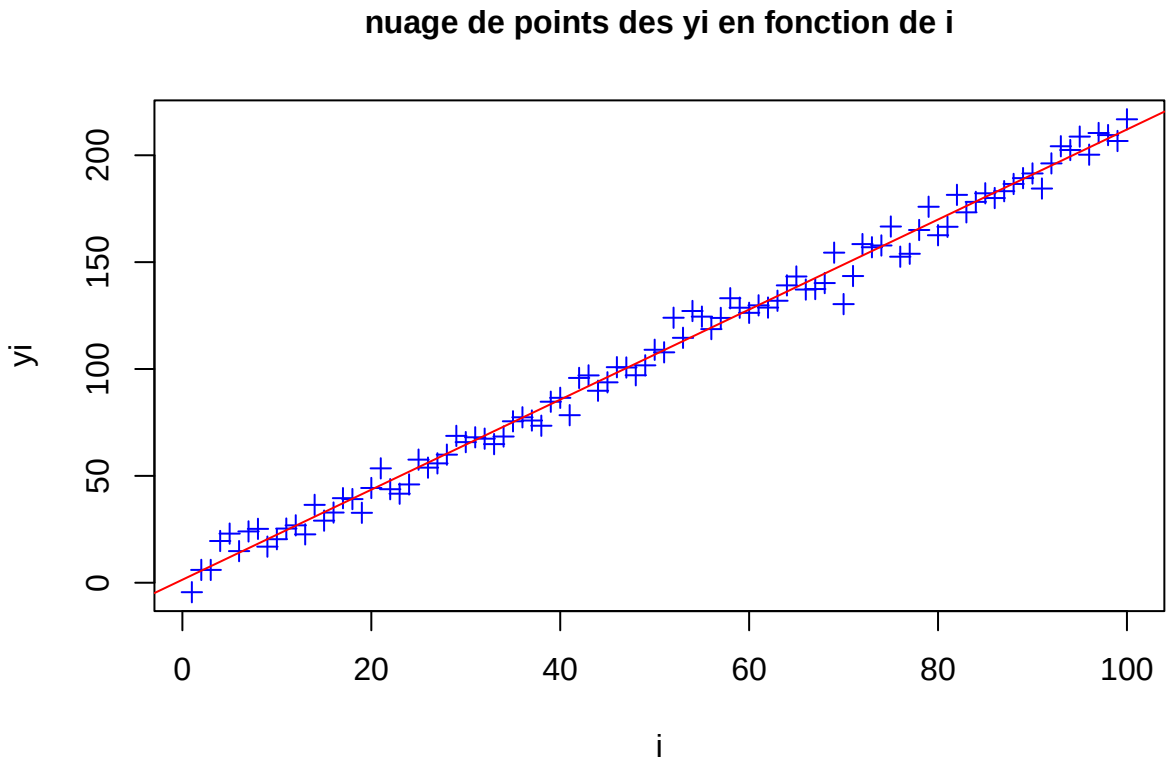

Figure 4: Nuage de points des yi en fonction de i

La figure 4 nous montre le nuage de points généré avec notre fonction, avec la droite de régression. Cette droite semble être une bon ajustement de notre fonction.

Exercice 18

On considère un tableau de contingence obtenu en ventilant 592 femmes suivant la couleur de leurs yeux et la couleur de leurs cheveux.

Table 17: Tableau de contingence du croisement entre couleur des yeux et des cheuveux

	blond	brun	chatin	roux
bleu	94	20	84	17
marron	7	68	119	26
noissette	10	15	54	14
vert	16	5	29	14

Le tableau 17 illustre ce tableau de contingence. On apprend par exemple que 94 femmes de nos données sont blonde au yeux bleu.

Table 18: matrice des fréquences du croisement entre couleur des yeux et des cheuveux en pourcent

	blond	brun	chatin	roux
bleu	16	3	14	3
marron	1	11	20	4
noissette	2	3	9	2
vert	3	1	5	2

Et voici la matrice des fréquences de nos données dans le tableau 18. On remarque que 11% des femmes de nos données sont brune au yeux marrons.

```
## Margins computed over dimensions
## in the following order:
## 1:
## 2:
```

Table 19: Tableau de contingence avec les marges

	blond	brun	chatin	roux	sum
bleu	94	20	84	17	215
marron	7	68	119	26	220
noissette	10	15	54	14	93
vert	16	5	29	14	64
sum	127	108	286	71	592

Nous pouvons ajouter les marges dans notre tableau de contingence comme le montre le tableau 19. Elles nous informe par exemple que 215 femmes de nos données on les yeux bleus, ou encore que 71 d'entre elles sont rousses.

```
## Margins computed over dimensions
## in the following order:
## 1:
## 2:
```

Table 20: Distribution conditionnelle des couleur des cheveux sachant la couleur des yeux

	blond	brun	chatin	roux	sum
bleu	43.72	9.30	39.07	7.91	100
marron	3.18	30.91	54.09	11.82	100
noissette	10.75	16.13	58.06	15.05	100
vert	25.00	7.81	45.31	21.88	100

Le tableau 20 des distribution conditionnelle de la couleur des cheveaux, sachant la couleur des yeux, nous apprend par exemple que, dans nos données, 58,06% des femmes au yeux noissettes on une couleur de cheveux chatin.

```
## Margins computed over dimensions
## in the following order:
## 1:
## 2:
```

Table 21: Distribution conditionnelle des couleurs des yeux sachant la couleur des cheveux

	blond	brun	chatin	roux
bleu	74.02	18.52	29.37	23.94
marron	5.51	62.96	41.61	36.62
noissette	7.87	13.89	18.88	19.72
vert	12.60	4.63	10.14	19.72
sum	100.00	100.00	100.00	100.00

Le tableau 21 correspond à l'inverse. Ici il s'agit de la distribution conditionnelle des couleur des yeux, sachant la couleur des cheveux. On apprend que les femmes brune de nos données ont pour 13.89% d'entre elles les yeux noissettes.

Table 22: Matrice taux de liaison

	blond	brun	chatin	roux
blond	1.000	-0.241	0.134	-0.155
brun	-0.241	1.000	0.920	0.986
chatin	0.134	0.920	1.000	0.918
roux	-0.155	0.986	0.918	1.000

Le tableau 22 nous apprend les laisions (variant entre -1 et 1), entre les modalité de la variable couleur de cheveaux. Par exemple on voit qu'il y a un lien positive entre la couleur chatin et brun, ce qui signifie qu'il y a une tendance similaire entre ces deux modalité vis à vis de la couleur des yeux.

```
##
##  Pearson's Chi-squared test
##
## data:  TEC
## X-squared = 138.29, df = 9, p-value < 2.2e-16
```

On remarque une pvaleur inferieur 2.2e-16, une valeur très proche de zero.Ce qui nous permet de rejeter l'hypothèse d'indépendance entre la couleur des yeux et celle des cheveaux.Il ya donc un lien à étudier entre ces deux variables.