Analyse exploratoire de données

Fabrice Rossi

Télécom ParisTech

Plan

Plan

Exploiter des données

- Que faire d'un paquet de données ?
- Comment exploiter le contenu d'un entrepôt de données ?

Exploiter des données

- Que faire d'un paquet de données ?
- Comment exploiter le contenu d'un entrepôt de données ?
- recensement
- 32561 personnes
- 15 attributs par personne

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Age	Workclass	Fnlwgt	Education	Educati on-num	Marital-status	Occupation	Relationship	Ethnicity	Gender	Capita l gain	Capit al loss	Hour per week	Native country	Salary
2	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
3	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
4	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
5	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
6	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
7	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
8	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
9	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
10	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
11	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
12	37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
13	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
14	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
15	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
16	40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
17	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
18	25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
19	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
20	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
21	43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
22	40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
23	54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
24	35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
25	43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
26	59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K
27	56	Local-gov	216851	Bachelors	13	Married-civ-spouse	Tech-support	Husband	White	Male	0	0	40	United-States	>50K
28	19	Private	168294	HS-grad	9	Never-married	Craft-repair	Own-child	White	Male	0	0	40	United-States	<=50K
29	54	?	180211	Some-college	10	Married-civ-spouse	?	Husband	Asian-Pac-Islander	Male	0	0	60	South	>50K
30	39	Private	367260	HS-grad	9	Divorced	Exec-managerial	Not-in-family	White	Male	0	0	80	United-States	<=50K
31	49	Private	193366	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States	<=50K
32	23	Local-gov	190709	Assoc-acdm	12	Never-married	Protective-serv	Not-in-family	White	Male	0	0	52	United-States	<=50K
33	20	Private	266015	Some-college	10	Never-married	Sales	Own-child	Black	Male	0	0	44	United-States	<=50K
34	45	Private	386940	Bachelors	13	Divorced	Exec-managerial	Own-child	White	Male	0	1408	40	United-States	<=50K

Exploiter des données

- Que faire d'un paquet de données ?
- Comment exploiter le contenu d'un entrepôt de données ?
- recensement
- 32561 personnes
- 15 attributs par personne

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Age	Workclass	Fnlwgt	Education	Educati on-num	Marital-status	Occupation	Relationship	Ethnicity	Gender	Capita l gain	Capit al loss	Hour per week	Native country	Salary
2	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
3	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
4	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
5	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
6	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
7	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
8	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
9	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
10	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
11	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
12	37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
13	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
14	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
15	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
16	40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
17	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
18	25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
19	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
20	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
21	43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
22	40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
23	54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
24	35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
25	43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
26	59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K
27	56	Local-gov	216851	Bachelors	13	Married-civ-spouse	Tech-support	Husband	White	Male	0	0	40	United-States	>50K
28	19	Private	168294	HS-grad	9	Never-married	Craft-repair	Own-child	White	Male	0	0	40	United-States	<=50K
29	54	?	180211	Some-college	10	Married-civ-spouse	?	Husband	Asian-Pac-Islander	Male	0	0	60	South	>50K
30	39	Private	367260	HS-grad	9	Divorced	Exec-managerial	Not-in-family	White	Male	0	0	80	United-States	<=50K
31	49	Private	193366	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States	<=50K
32	23	Local-gov	190709	Assoc-acdm	12	Never-married	Protective-serv	Not-in-family	White	Male	0	0	52	United-States	<=50K
33	20	Private	266015	Some-college	10	Never-married	Sales	Own-child	Black	Male	0	0	44	United-States	<=50K
34	45	Private	386940	Bachelors	13	Divorced	Exec-managerial	Own-child	White	Male	0	1408	40	United-States	<=50K

- Volume classique : milliers à millions de lignes, dizaine à centaines de colonnes
- Exploration systématique impossible (même pour de petits paquets de données)

Outils d'exploitation

- Support informatique et mathématique :
  - outils d'exploitation des données
  - but : diminuer la charge cognitive pour l'analyste
- Deux grandes classes d'outils :
  1. exploration
  2. modélisation

Outils d'exploitation

- Support informatique et mathématique :
  - outils d'exploitation des données
  - but : diminuer la charge cognitive pour l'analyste
- Deux grandes classes d'outils :
  1. exploration
     - pas d'idée *a priori* sur les données
     - recherche de régularité (dépendances, groupes homogènes, etc.)
  2. modélisation

Outils d'exploitation

- Support informatique et mathématique :
  - outils d'exploitation des données
  - but : diminuer la charge cognitive pour l'analyste
- Deux grandes classes d'outils :
  1. exploration
     - pas d'idée *a priori* sur les données
     - recherche de régularité (dépendances, groupes homogènes, etc.)
  2. modélisation
     - idée précise sur les données
     - construction de modèles prédictifs

Outils d'exploitation

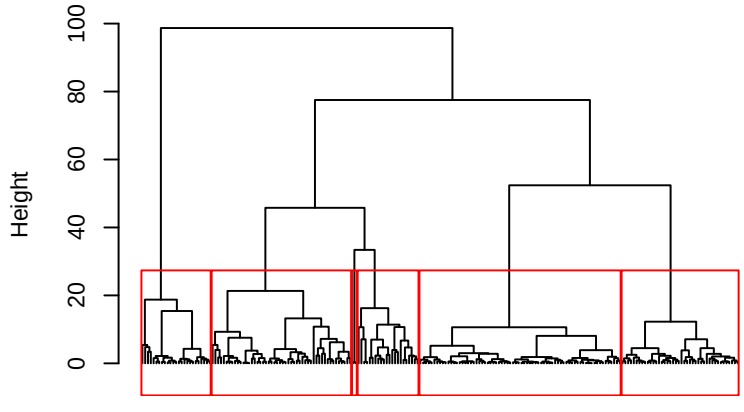
- Support informatique et mathématique :
  - outils d'exploitation des données
  - but : diminuer la charge cognitive pour l'analyste
- Deux grandes classes d'outils :
  1. exploration
     - pas d'idée *a priori* sur les données
     - recherche de régularité (dépendances, groupes homogènes, etc.)
  2. modélisation
     - idée précise sur les données
     - construction de modèles prédictifs
- outil utilisé : R (`http://R-project.org/`)

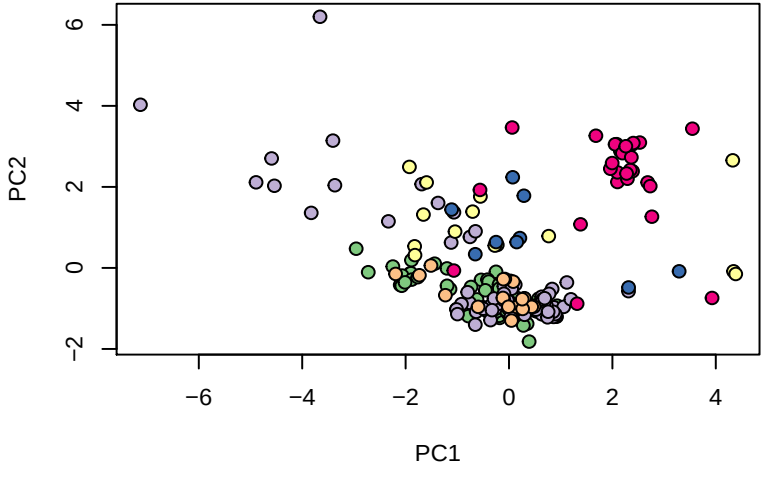
Analyse exploratoire

- Objectifs :
  - obtenir une vision globale d'un jeu de données
  - découvrir des formes de régularité
- Moyens :
  - représentations visuelles (et interactives) des données
  - recherche automatique de régularités :
    - corrélation et dépendance entre variables
    - groupes homogènes (classification)
    - schémas fréquents

Analyse exploratoire

- Objectifs :
  - obtenir une vision globale d'un jeu de données
  - découvrir des formes de régularité
- Moyens :
  - représentations visuelles (et interactives) des données
  - recherche automatique de régularités :
    - corrélation et dépendance entre variables
    - groupes homogènes (classification)
    - schémas fréquents





Modélisation

- Objectifs :
  - inférer des informations inconnues
  - prédire l'évolution des données
- Moyens :
  - données d'apprentissage :
    - connaître l'évolution d'une grandeur dans le passé pour prédire son évolution future (données historiques)
    - connaître une propriété de certains objets (par exemple le salaire de certains clients) pour inférer sa valeur pour les autres objets
  - méthodes d'apprentissage : construire un modèle à partir des données d'apprentissage

Modélisation

- Objectifs :
  - inférer des informations inconnues
  - prédire l'évolution des données
- Moyens :
  - données d'apprentissage :
    - connaître l'évolution d'une grandeur dans le passé pour prédire son évolution future (données historiques)
    - connaître une propriété de certains objets (par exemple le salaire de certains clients) pour inférer sa valeur pour les autres objets
  - méthodes d'apprentissage : construire un modèle à partir des données d'apprentissage
- Stratégie :
  - analyse exploratoire
  - formulation d'hypothèses
  - construction de modèles pour valider les hypothèses

Modèle mathématique

- On a N observations, les $z_i \in \mathcal{Z}$
- Modèle statistique/probabiliste
  - il existe une distribution P_Z sur $\mathcal{Z}$ inconnue
  - les z_i sont des réalisations de variables aléatoires avec cette distribution
  - les variables aléatoires sont indépendantes (en général)

Modèle mathématique

- On a N observations, les $z_i \in \mathcal{Z}$
- Modèle statistique/probabiliste
  - il existe une distribution P_Z sur $\mathcal{Z}$ inconnue
  - les z_i sont des réalisations de variables aléatoires avec cette distribution
  - les variables aléatoires sont indépendantes (en général)
- En général
  - $\mathcal{Z} = \Pi_{p=1}^{P} \mathcal{Z}_p$: P variables pour décrire chaque objet
  - quand $\mathcal{Z}_p \subset \mathbb{R}$: variable numérique (ou ordonnée)
  - quand $\mathcal{Z}_p = \{a, b, \ldots\}$: variable nominale (un nombre fini de valeurs possibles non ordonnées)

Plan

Analyses élémentaires

- Première étape d'une analyse exploratoire
  - travailler variable par variable
  - numériquement et graphiquement
- Variable numérique
  - à valeurs dans $\mathbb{R}$
  - statistiques classiques : moyenne, variance, médiane, etc.
  - représentations associées : histogramme, *boxplot*

Analyses élémentaires

- Première étape d'une analyse exploratoire
  - travailler variable par variable
  - numériquement et graphiquement
- Variable numérique
  - à valeurs dans $\mathbb{R}$
  - statistiques classiques : moyenne, variance, médiane, etc.
  - représentations associées : histogramme, *boxplot*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Age	Workclass	Fnlwgt	Education	Education-num	Marital-status	Occupation	Relationship	Ethnicity	Gender	Capital gain	Capital loss	Hour per week	Native country	Salary
2	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
3	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
4	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
5	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
6	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
7	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
8	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
9	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
10	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
11	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
12	37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
13	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
14	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
15	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
16	40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
17	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
18	25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
19	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
20	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
21	43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
22	40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
23	54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
24	35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
25	43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
26	59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K
27	56	Local-gov	216851	Bachelors	13	Married-civ-spouse	Tech-support	Husband	White	Male	0	0	40	United-States	>50K
28	19	Private	168294	HS-grad	9	Never-married	Craft-repair	Own-child	White	Male	0	0	40	United-States	<=50K
29	54	?	180211	Some-college	10	Married-civ-spouse	?	Husband	Asian-Pac-Islander	Male	0	0	60	South	>50K
30	39	Private	367260	HS-grad	9	Divorced	Exec-managerial	Not-in-family	White	Male	0	0	80	United-States	<=50K
31	49	Private	193366	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States	<=50K
32	23	Local-gov	190709	Assoc-acdm	12	Never-married	Protective-serv	Not-in-family	White	Male	0	0	52	United-States	<=50K
33	20	Private	266015	Some-college	10	Never-married	Sales	Own-child	Black	Male	0	0	44	United-States	<=50K
34	45	Private	386940	Bachelors	13	Divorced	Exec-managerial	Own-child	White	Male	0	1408	40	United-States	<=50K

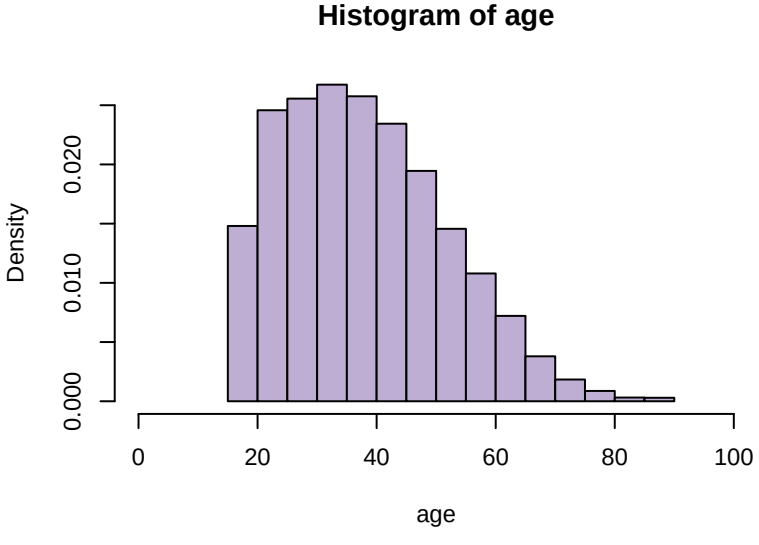
Variable âge : numérique

Analyses élémentaires

- Première étape d'une analyse exploratoire
  - travailler variable par variable
  - numériquement et graphiquement
- Variable numérique
  - à valeurs dans $\mathbb{R}$
  - statistiques classiques : moyenne, variance, médiane, etc.
  - représentations associées : histogramme, *boxplot*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Age	Workclass	Fnlwgt	Education	Education-num	Marital-status	Occupation	Relationship	Ethnicity	Gender	Capital gain	Capital loss	Hour per week	Native country	Salary
2	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
3	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
4	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
5	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
6	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
7	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
8	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16	Jamaica	<=50K
9	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
10	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
11	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
12	37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
13	30	State-gov	141297	Bachelors	13	Married-civ-spouse	Prof-specialty	Husband	Asian-Pac-Islander	Male	0	0	40	India	>50K
14	23	Private	122272	Bachelors	13	Never-married	Adm-clerical	Own-child	White	Female	0	0	30	United-States	<=50K
15	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
16	40	Private	121772	Assoc-voc	11	Married-civ-spouse	Craft-repair	Husband	Asian-Pac-Islander	Male	0	0	40	?	>50K
17	34	Private	245487	7th-8th	4	Married-civ-spouse	Transport-moving	Husband	Amer-Indian-Eskimo	Male	0	0	45	Mexico	<=50K
18	25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
19	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
20	38	Private	28887	11th	7	Married-civ-spouse	Sales	Husband	White	Male	0	0	50	United-States	<=50K
21	43	Self-emp-not-inc	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
22	40	Private	193524	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	60	United-States	>50K
23	54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20	United-States	<=50K
24	35	Federal-gov	76845	9th	5	Married-civ-spouse	Farming-fishing	Husband	Black	Male	0	0	40	United-States	<=50K
25	43	Private	117037	11th	7	Married-civ-spouse	Transport-moving	Husband	White	Male	0	2042	40	United-States	<=50K
26	59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40	United-States	<=50K
27	56	Local-gov	216851	Bachelors	13	Married-civ-spouse	Tech-support	Husband	White	Male	0	0	40	United-States	>50K
28	19	Private	168294	HS-grad	9	Never-married	Craft-repair	Own-child	White	Male	0	0	40	United-States	<=50K
29	54	?	180211	Some-college	10	Married-civ-spouse	?	Husband	Asian-Pac-Islander	Male	0	0	60	South	>50K
30	39	Private	367260	HS-grad	9	Divorced	Exec-managerial	Not-in-family	White	Male	0	0	80	United-States	<=50K
31	49	Private	193366	HS-grad	9	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	40	United-States	<=50K
32	23	Local-gov	190709	Assoc-acdm	12	Never-married	Protective-serv	Not-in-family	White	Male	0	0	52	United-States	<=50K
33	20	Private	266015	Some-college	10	Never-married	Sales	Own-child	Black	Male	0	0	44	United-States	<=50K
34	45	Private	386940	Bachelors	13	Divorced	Exec-managerial	Own-child	White	Male	0	1408	40	United-States	<=50K

Variable âge : numérique

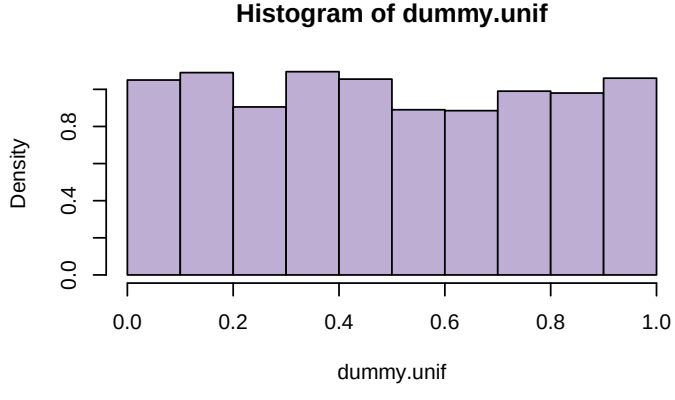


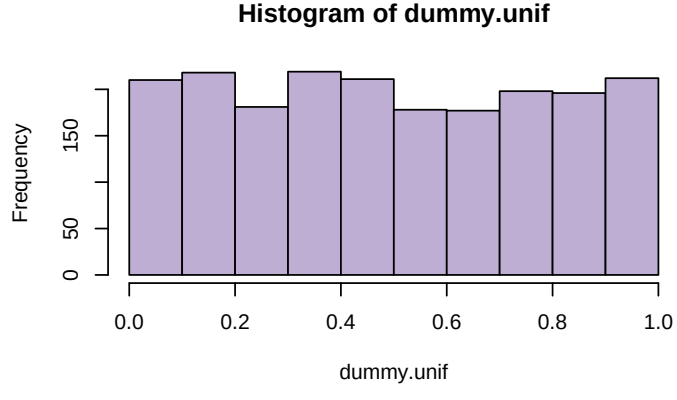
Histogramme

- Un histogramme représente une estimation de la distribution d'une variable
- Principe de construction :
  - division de l'intervalle $[\min, \max]$ en K sous-intervalles (diverses règles pour K, par exemple $\sim \log N$)
  - dénombrement des objets pour lesquels la valeur de la variable tombe dans chacun des intervalles
  - représentation par des barres de surfaces proportionnelles aux décomptes

Histogramme

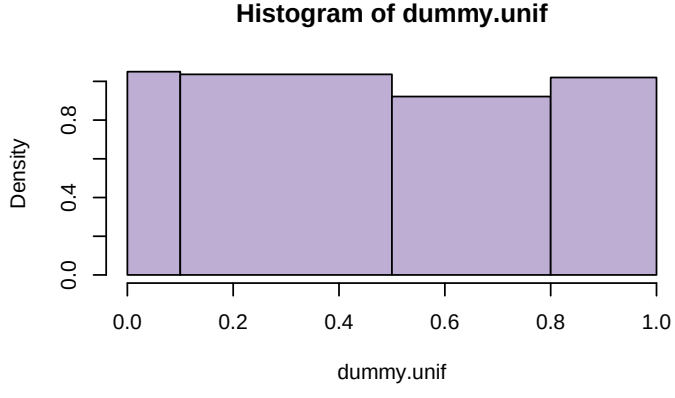
- Un histogramme représente une estimation de la distribution d'une variable
- Principe de construction :
  - division de l'intervalle $[\min, \max]$ en K sous-intervalles (diverses règles pour K, par exemple $\sim \log N$)
  - dénombrement des objets pour lesquels la valeur de la variable tombe dans chacun des intervalles
  - représentation par des barres de surfaces proportionnelles aux décomptes
- Attention aux intervalles de longueurs différentes

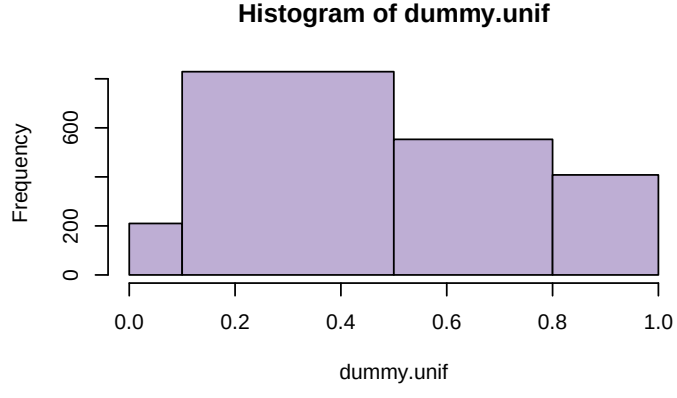




Histogramme

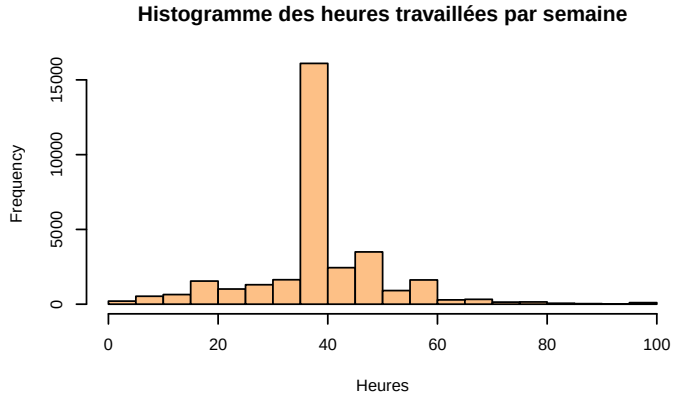
- Un histogramme représente une estimation de la distribution d'une variable
- Principe de construction :
  - division de l'intervalle [min, max] en K sous-intervalles (diverses règles pour K, par exemple $\sim \log N$)
  - dénombrement des objets pour lesquels la valeur de la variable tombe dans chacun des intervalles
  - représentation par des barres de surfaces proportionnelles aux décomptes
- Attention aux intervalles de longueurs différentes



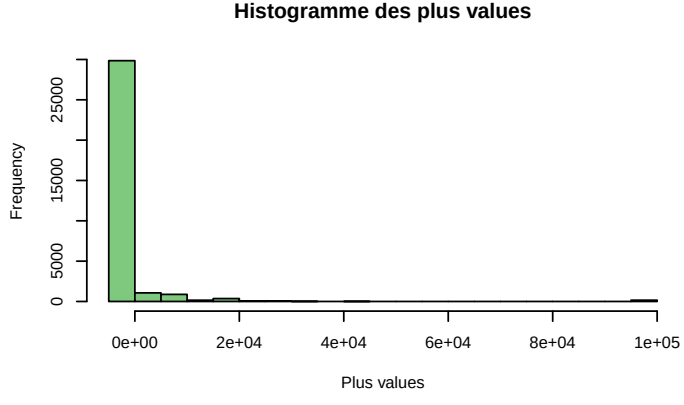


Intérêts

Âge



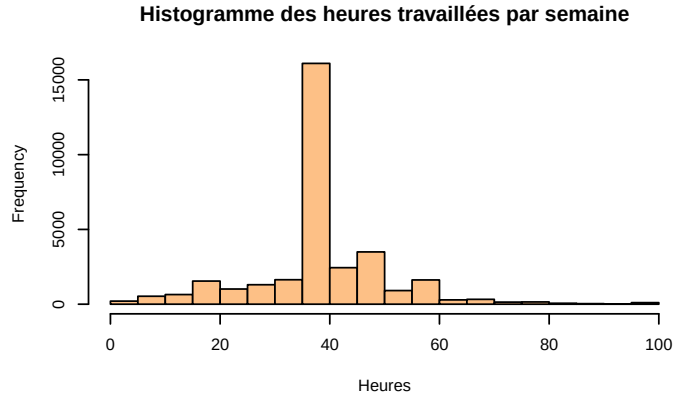
Temps de travail



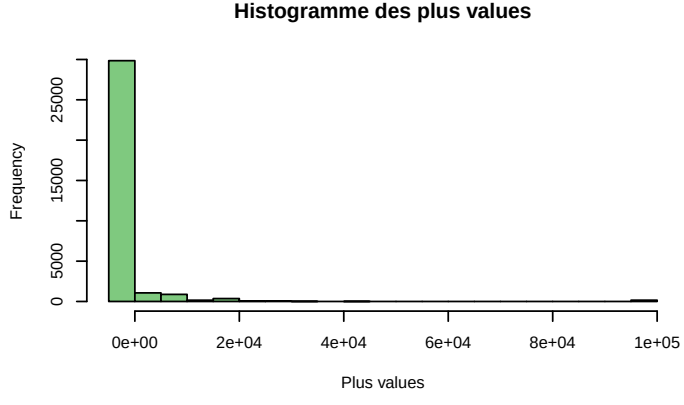
Plus values

Intérêts

Âge



Temps de travail

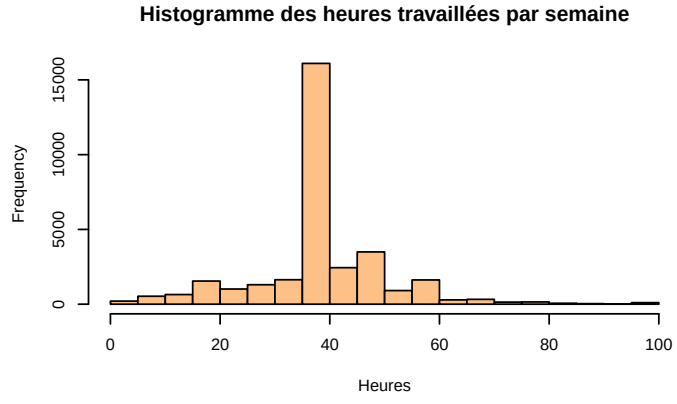


Plus values

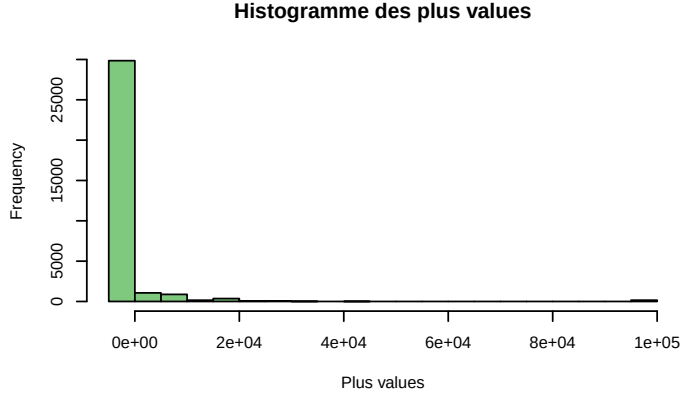
- Idée générale de la distribution

Intérêts

Âge



Temps de travail

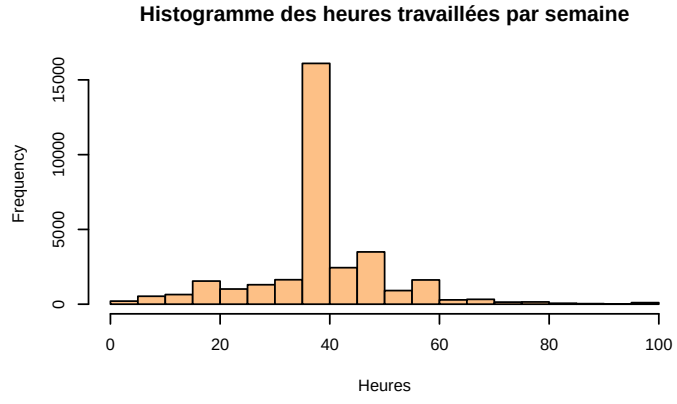


Plus values

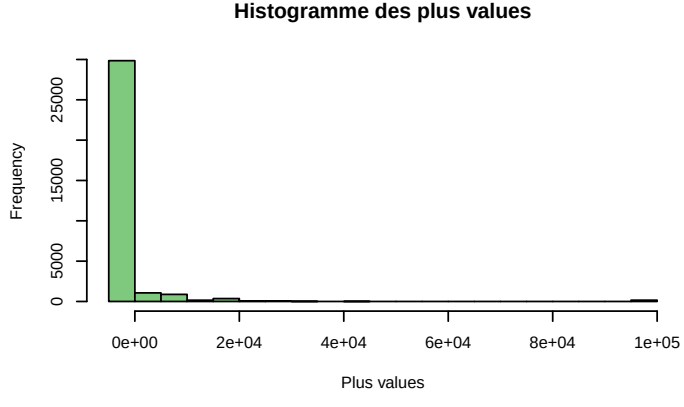
- Idée générale de la distribution
- “irrégularités”

Intérêts

Âge



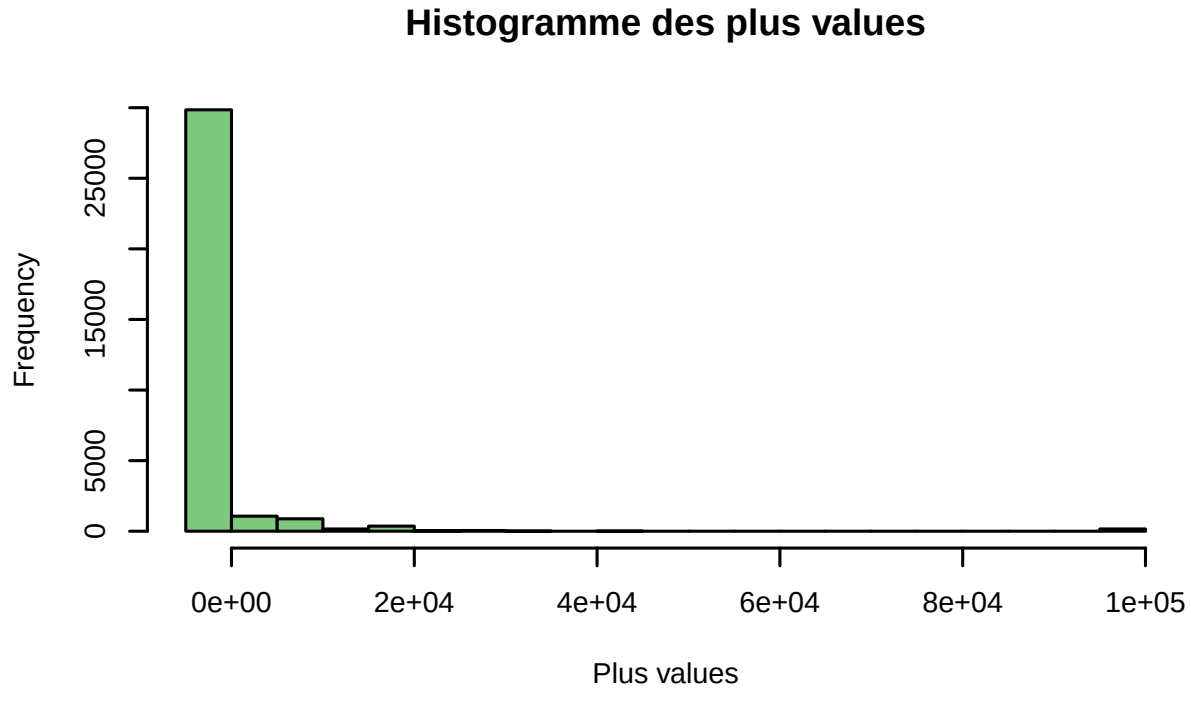
Temps de travail



Plus values

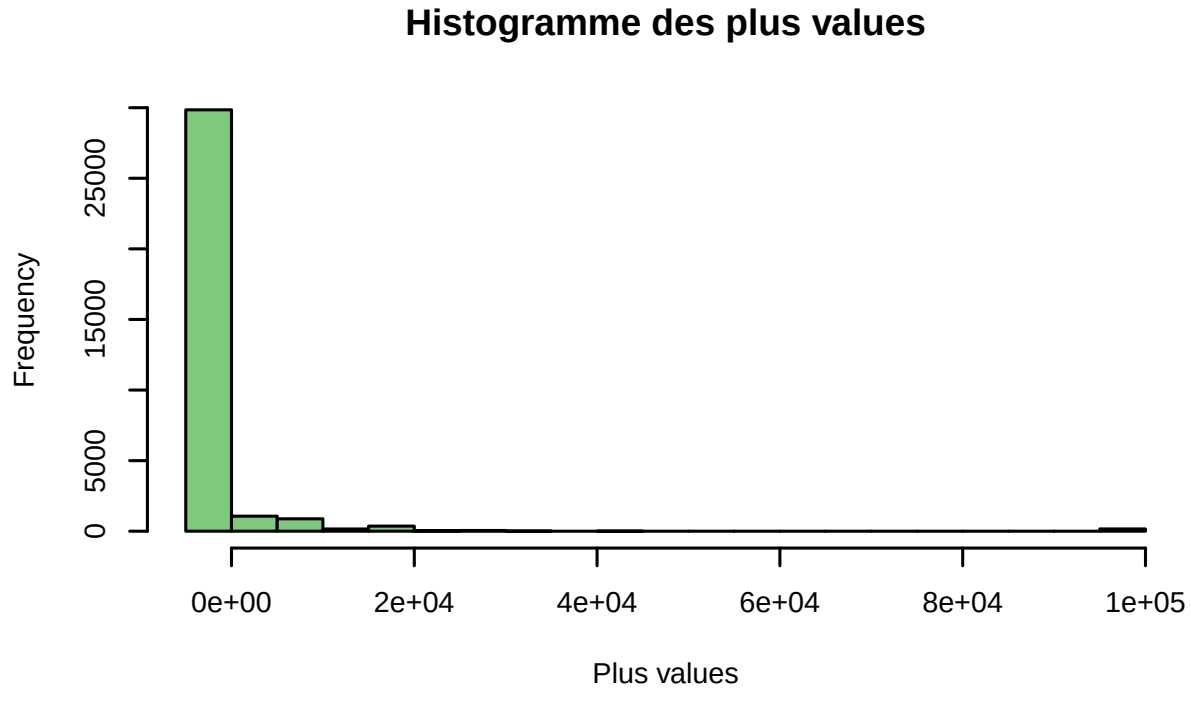
- Idée générale de la distribution
- “irrégularités”
- distribution complètement atypique

Limitations



- presque aucune information :
  - presque toutes les valeurs sont négatives
  - quelques valeurs très grandes

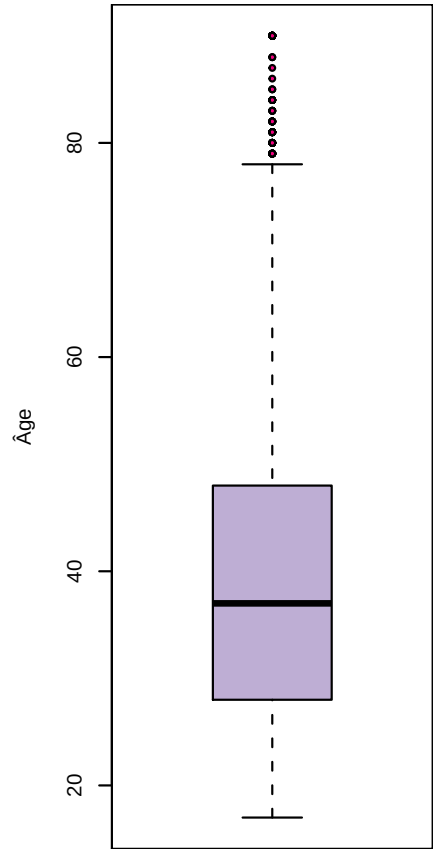
Limitations



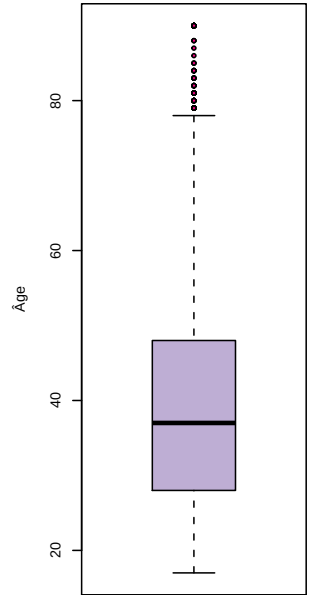
- presque aucune information :
  - presque toutes les valeurs sont négatives
  - quelques valeurs très grandes
- comparaisons difficiles (cf la suite)

Boxplot

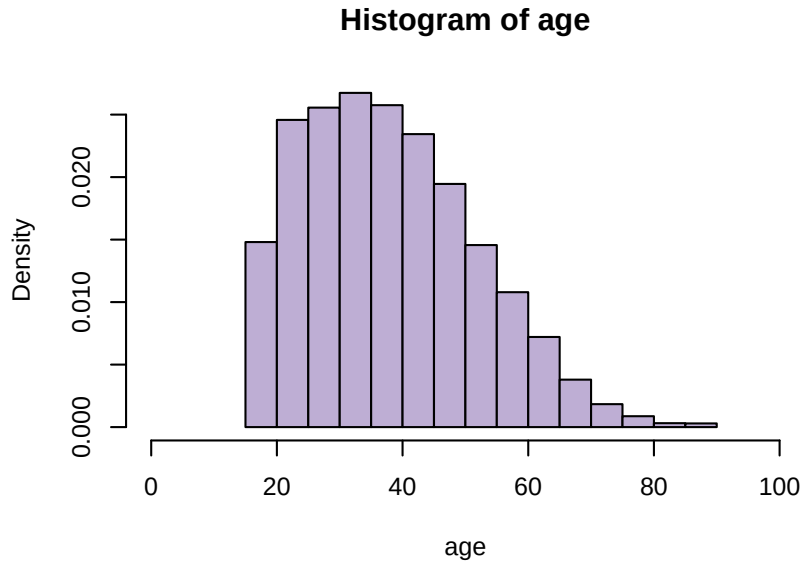
- a.k.a. boîte à moustaches ou boîte à pattes
- Représentation compacte d'une distribution
  - ligne centrale : médiane
  - ligne basse : premier quartile
  - ligne haute : troisième quartile
  - moustaches :
    - le max du min et de la médiane - 1.5 l'intervalle interquartile
    - le min du max et de la médiane + 1.5 l'intervalle interquartile
  - points atypiques (*outliers*) : au delà des moustaches

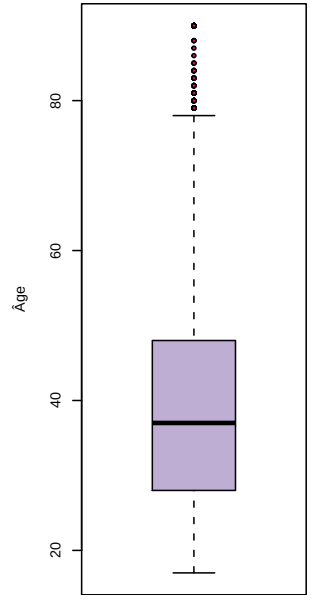


Comparaison



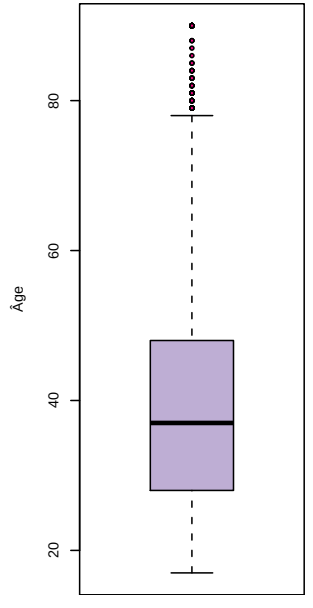
Comparaison





- plus d'information
- plus dépouillé

Comparaison



- plus d'information
- inférence moins précise

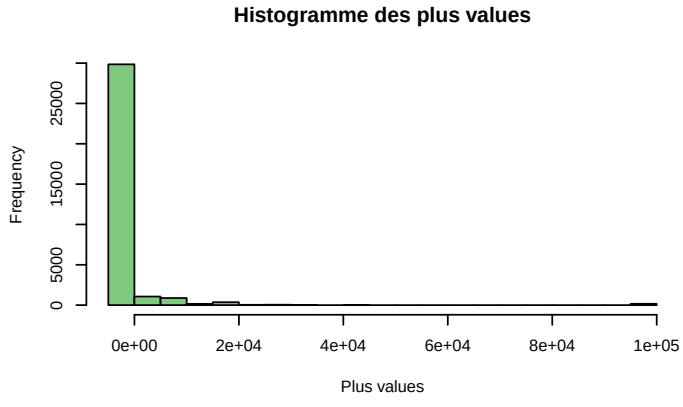
- plus dépouillé
- quelques informations très précises

Statistiques

- Indicateurs classiques :
  - tendance : moyenne et médiane
  - dispersion : écart-type, intervalle interquartile

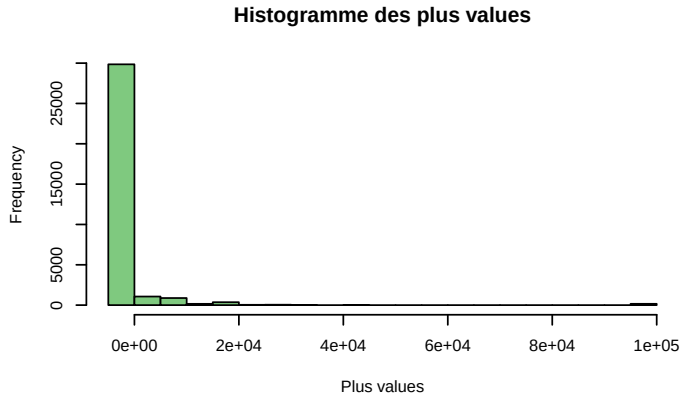
Statistiques

- Indicateurs classiques :
  - tendance : moyenne et médiane
  - dispersion : écart-type, intervalle interquartile
- Interprétation parfois délicate :
  - moyenne $= 990$
  - médiane $= 0$
  - écart-type $= 7410$
  - intervalle interquartile $= 0$



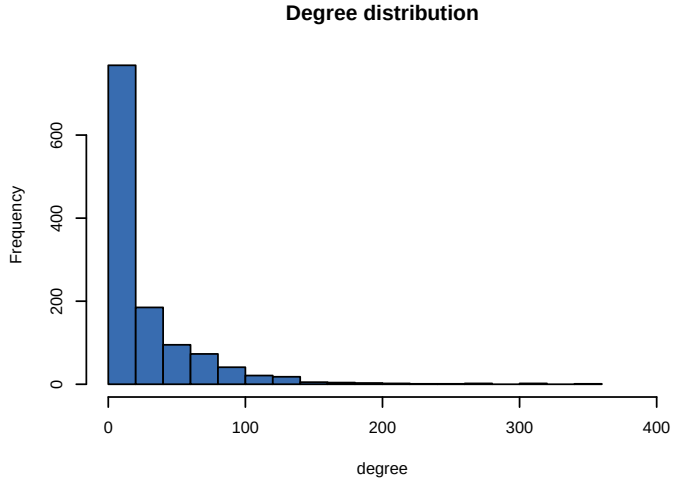
Statistiques

- Indicateurs classiques :
  - tendance : moyenne et médiane
  - dispersion : écart-type, intervalle interquartile
- Interprétation parfois délicate :
  - moyenne $= 990$
  - médiane $= 0$
  - écart-type $= 7410$
  - intervalle interquartile $= 0$
  - meilleurs choix ici :
    - 87 % des personnes ont une plus value nulle, 8.3 % positive et 4.7 % négative
    - puis statistiques sur les deux groupes (par ex., perte médiane 1887)



Sens des statistiques

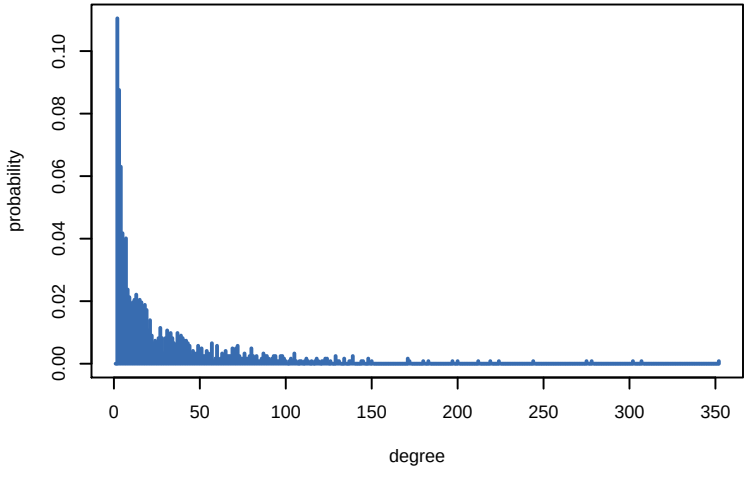
- La pertinence de la statistique dépend de la distribution
- Exemple :
  - blogs politiques
  - graphe des liens entre les blogs (blogroll)
  - distribution des degrés des noeuds

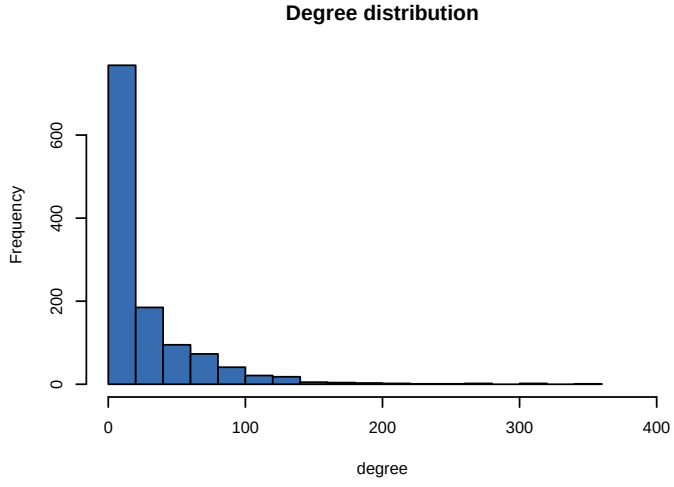


- $\mu = 27.36$, $\sigma = 38.42$
- $m = 13$, $\delta = 33$

Sens des statistiques

- La pertinence de la statistique dépend de la distribution
- Exemple :
  - blogs politiques
  - graphe des liens entre les blogs (blogroll)
  - distribution des degrés des noeuds

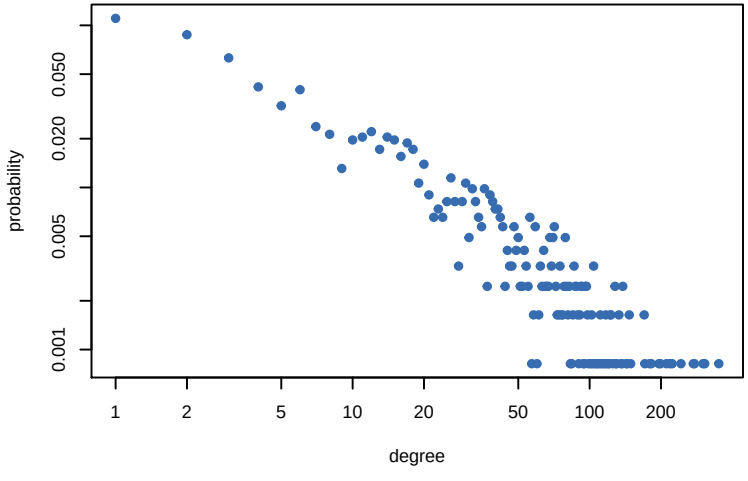


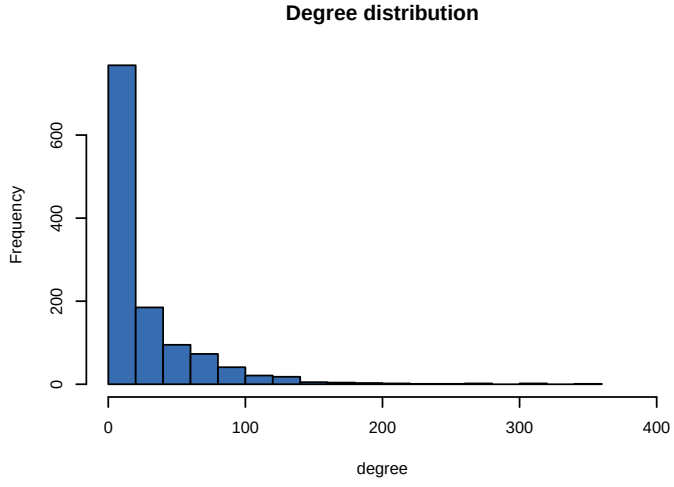


- $\mu = 27.36$, $\sigma = 38.42$
- $m = 13$, $\delta = 33$
- loi puissance : $P(x) \simeq x^{-\alpha}$

Sens des statistiques

- La pertinence de la statistique dépend de la distribution
- Exemple :
  - blogs politiques
  - graphe des liens entre les blogs (blogroll)
  - distribution des degrés des noeuds

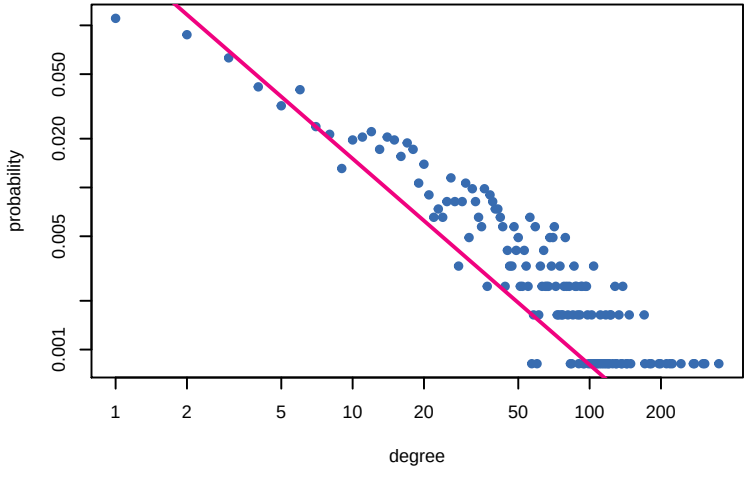


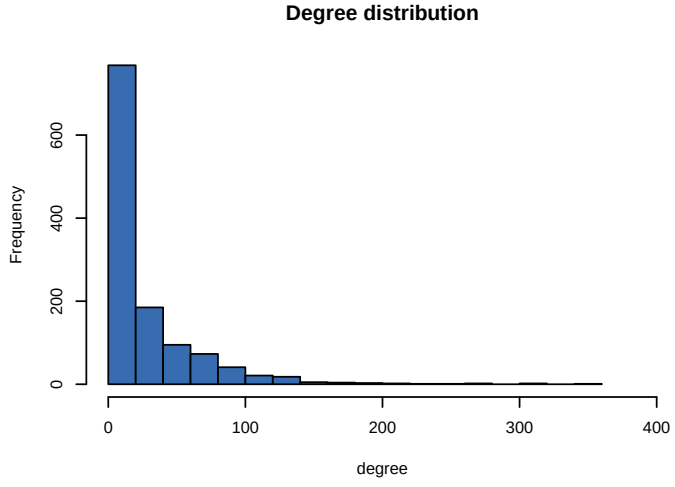


- $\mu = 27.36$, $\sigma = 38.42$
- $m = 13$, $\delta = 33$
- loi puissance : $P(x) \simeq x^{-\alpha}$
- sans échelle : la moyenne informe peu

Sens des statistiques

- La pertinence de la statistique dépend de la distribution
- Exemple :
  - blogs politiques
  - graphe des liens entre les blogs (blogroll)
  - distribution des degrés des noeuds

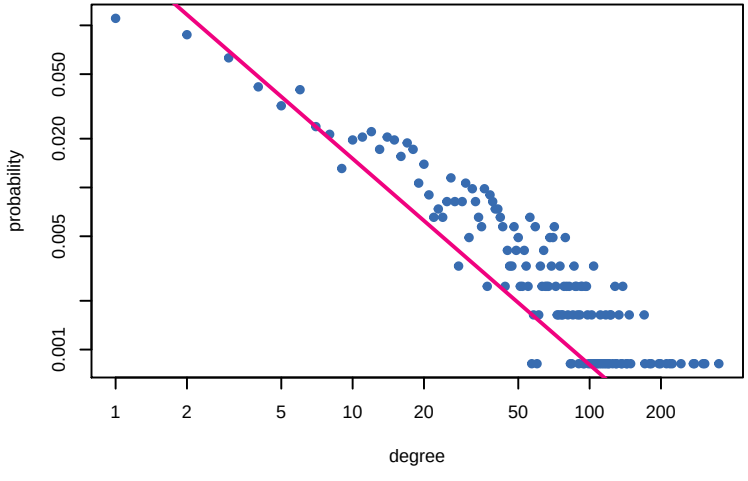


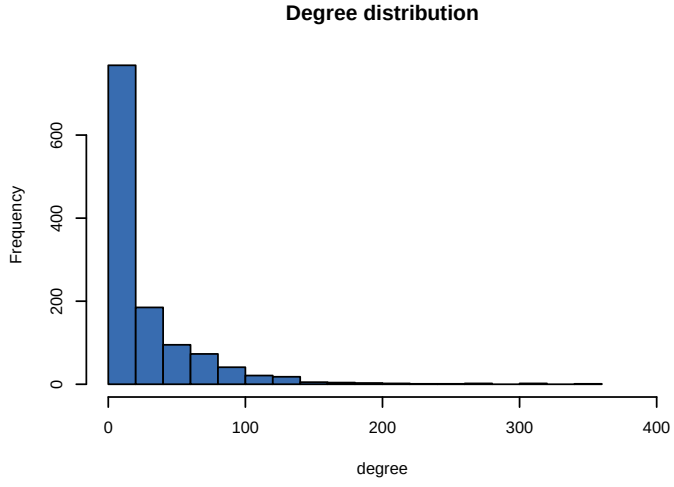


- $\mu = 27.36$, $\sigma = 38.42$
- $m = 13$, $\delta = 33$
- loi puissance : $P(x) \simeq x^{-\alpha}$
- sans échelle : la moyenne informe peu
- ici $\alpha \simeq 1.27$

Sens des statistiques

- La pertinence de la statistique dépend de la distribution
- Exemple :
  - blogs politiques
  - graphe des liens entre les blogs (blogroll)
  - distribution des degrés des noeuds

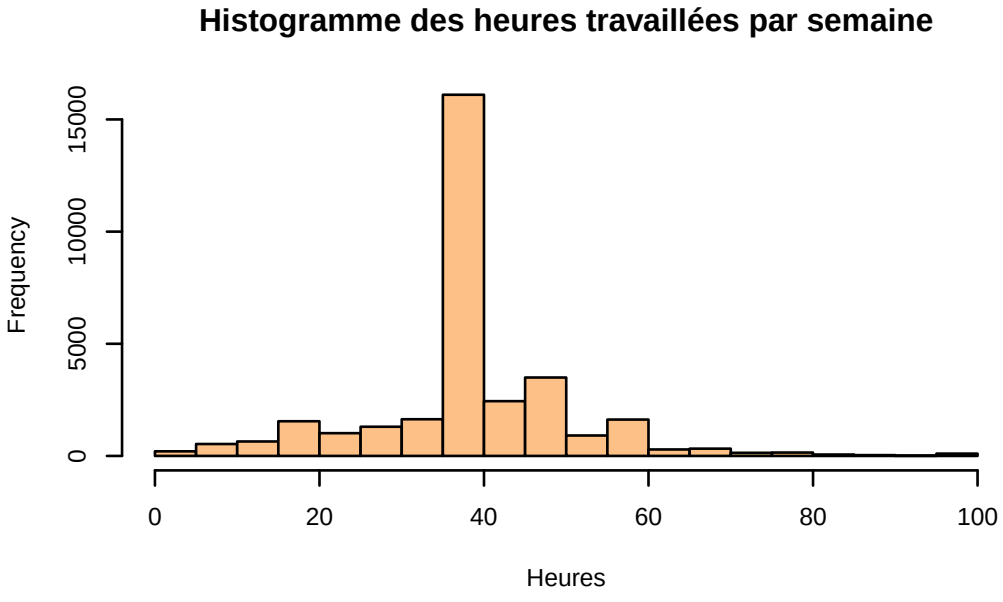


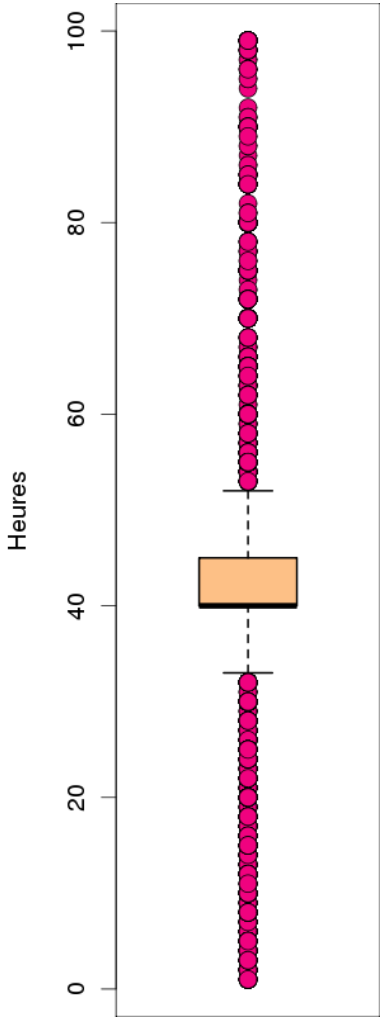


- $\mu = 27.36$, $\sigma = 38.42$
- $m = 13$, $\delta = 33$
- loi puissance : $P(x) \simeq x^{-\alpha}$
- sans échelle : la moyenne informe peu
- ici $\alpha \simeq 1.27$

Adapter les statistiques aux données

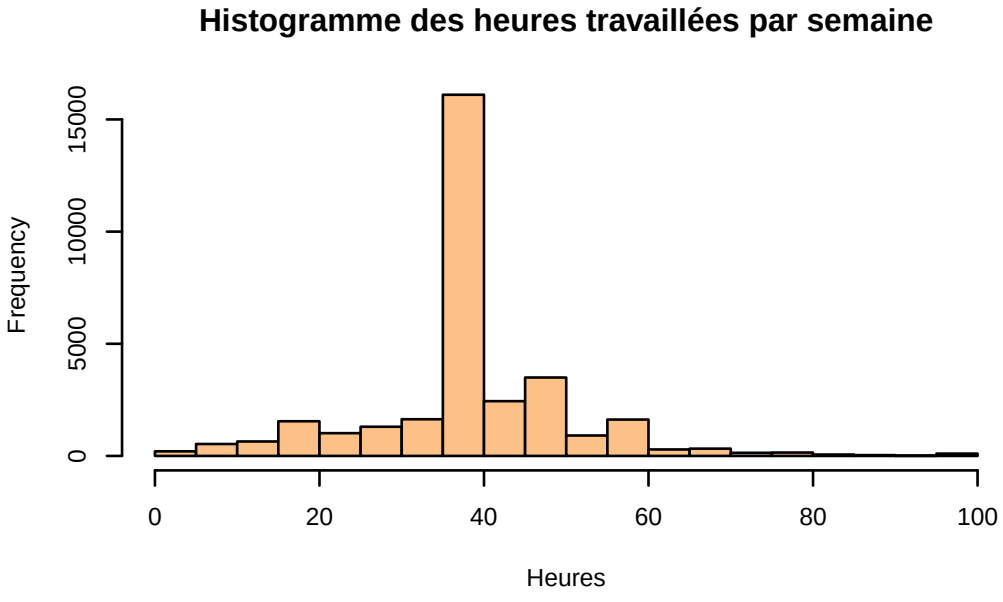
Trois points de vue

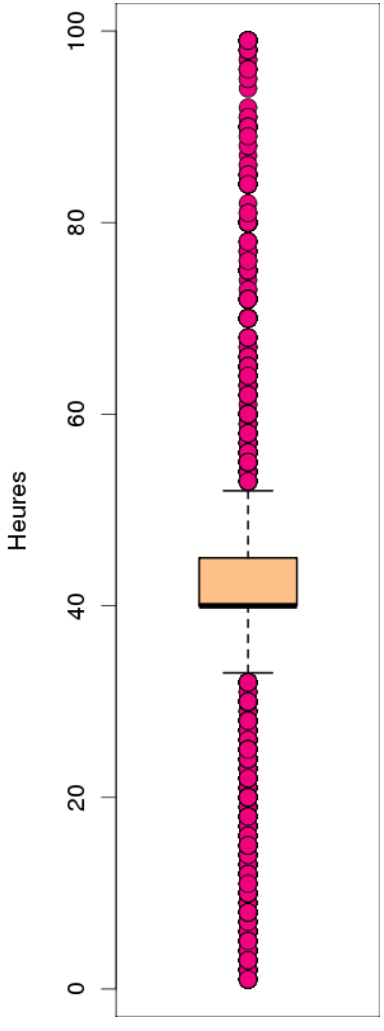




- Moyenne : 40.44, Écart-type : 12.35
- Médiane : 40, Interquartile : 5

Trois points de vue





- Moyenne : 40.44, Écart-type : 12.35
- Médiane : 40, Interquartile : 5
- Compléments :
  - 47 % = 40 heures
  - 29 % > 40 heures
  - 24 % < 40 heures

Variables nominales

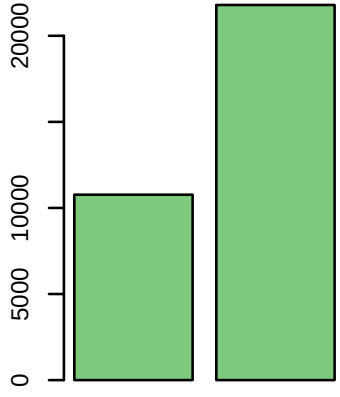
- variable nominale (ou qualitative) : variable à valeurs dans un ensemble fini quelconque (les modalités)
- quand les modalités sont ordonnées : variable ordinale

Variables nominales

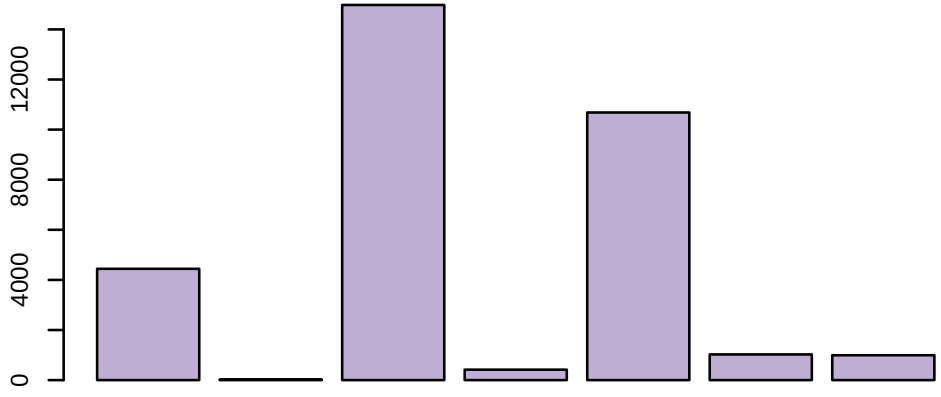
- variable nominale (ou qualitative) : variable à valeurs dans un ensemble fini quelconque (les modalités)
- quand les modalités sont ordonnées : variable ordinale
- représentation par un diagramme à bâtons :
  - un bâton par modalité
  - hauteur proportionnelle à la fréquence de la modalité
  - ordre arbitraire sauf dans la cas ordinal

Variables nominales

- variable nominale (ou qualitative) : variable à valeurs dans un ensemble fini quelconque (les modalités)
- quand les modalités sont ordonnées : variable ordinale
- représentation par un diagramme à bâtons :
  - un bâton par modalité
  - hauteur proportionnelle à la fréquence de la modalité
  - ordre arbitraire sauf dans la cas ordinal



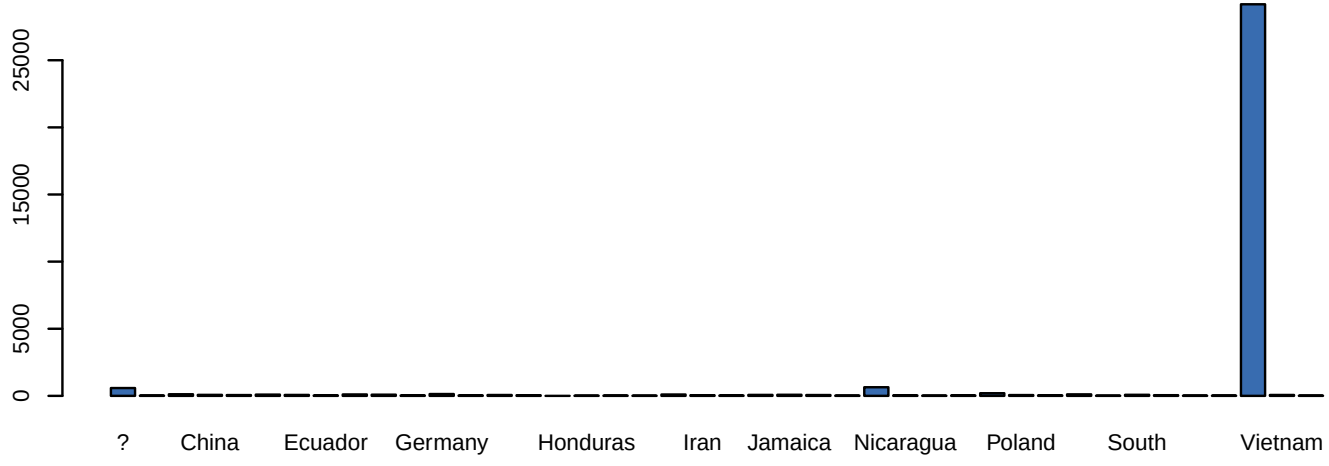
Genre



Statut marital

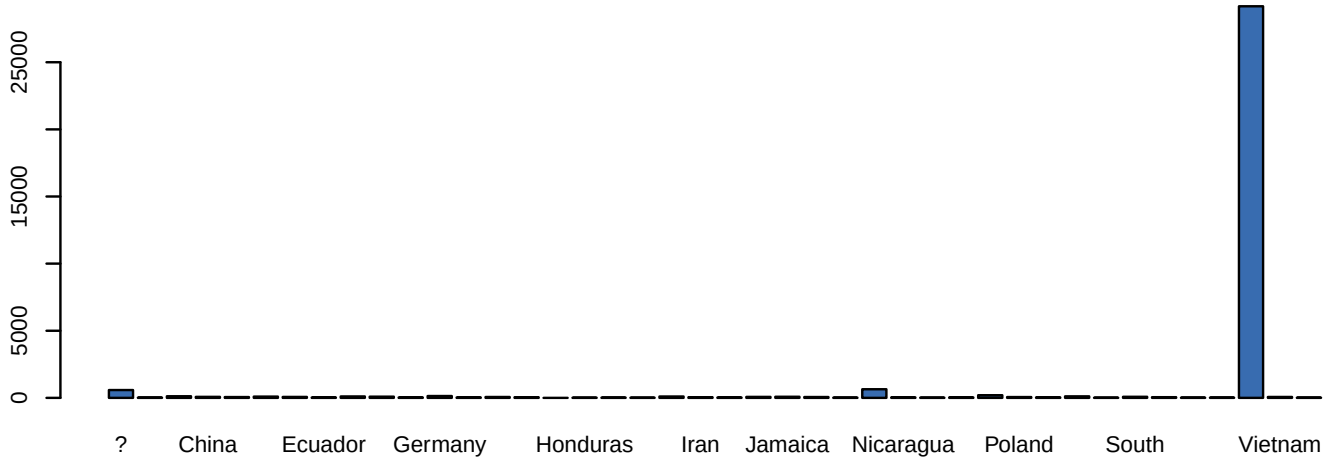
Lisibilité

- Déséquilibre

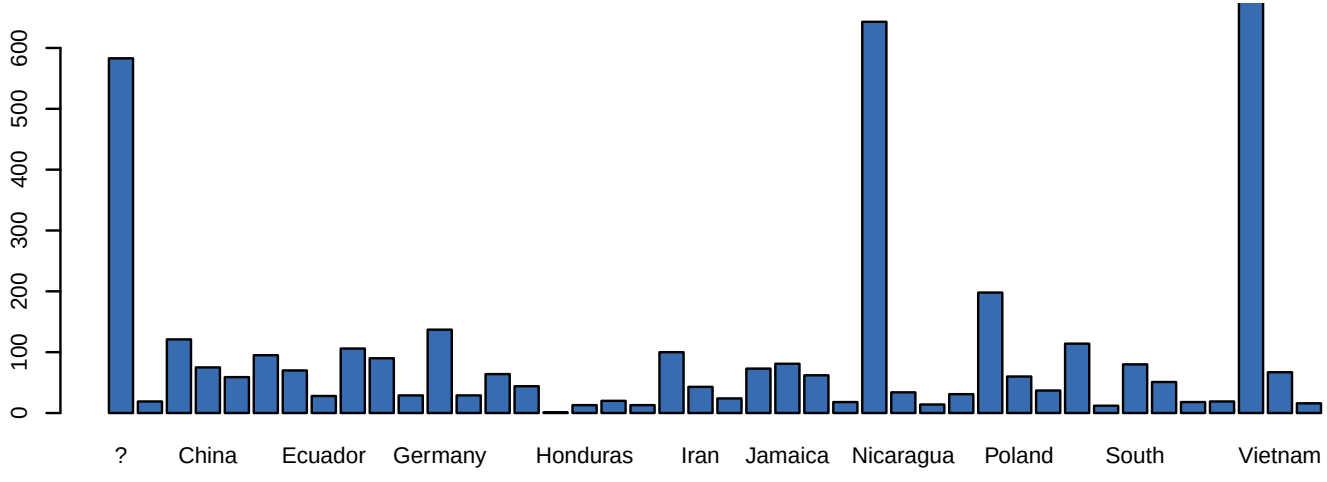


Lisibilité

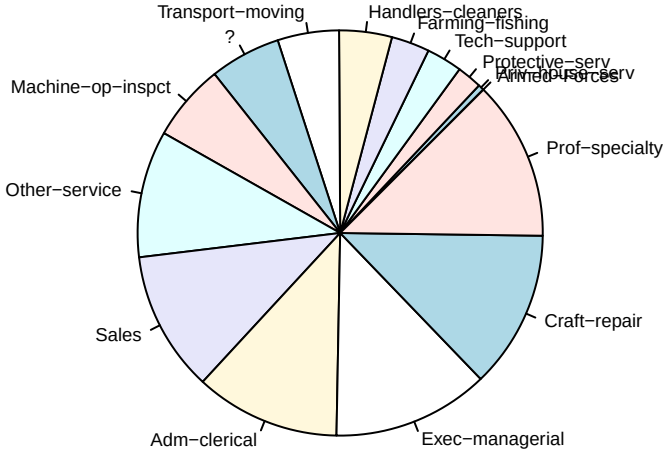
- Déséquilibre



- Grand nombre de modalités

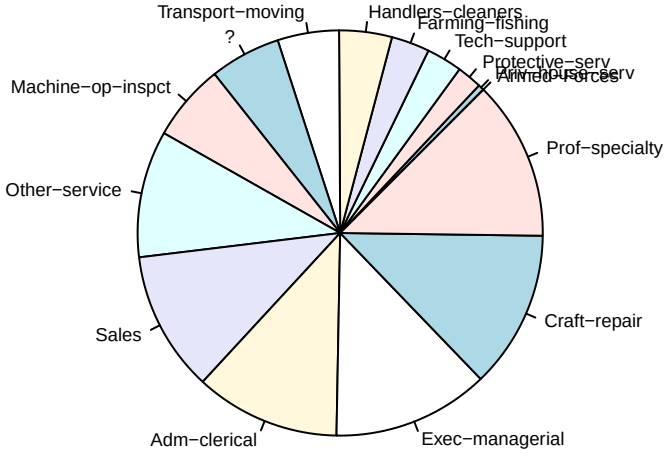


Camembert

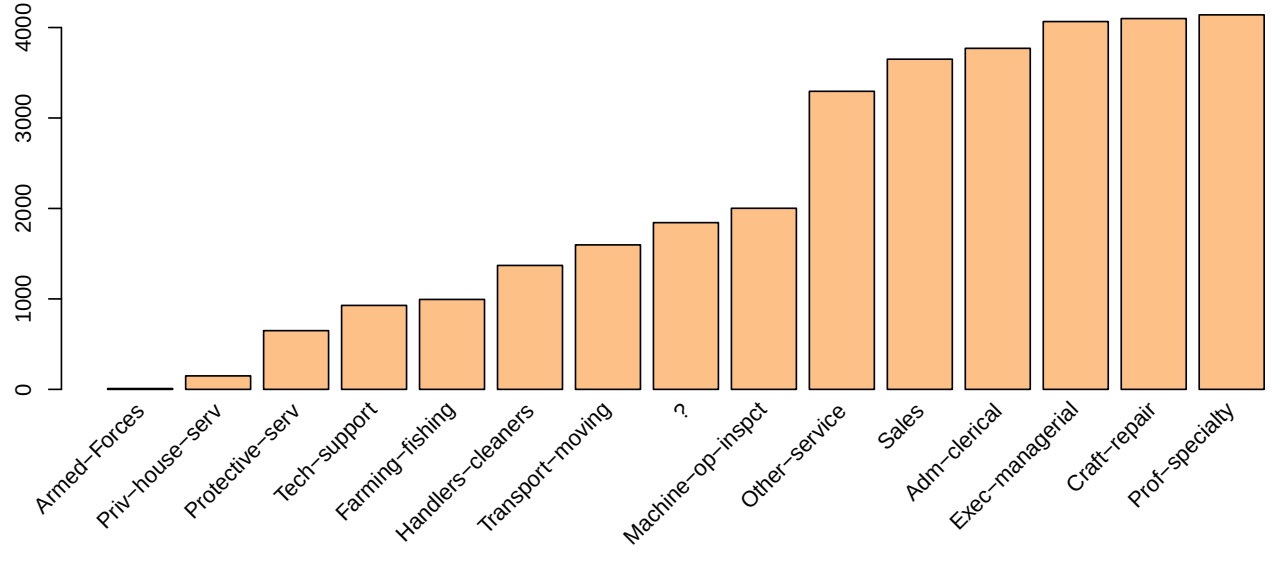


- représentation très classique
- versions “créatives” (3D...)
- mauvaise solution : lecture des surfaces et des angles difficiles

Camembert



- représentation très classique
- versions “créatives” (3D...)
- mauvaise solution : lecture des surfaces et des angles difficiles



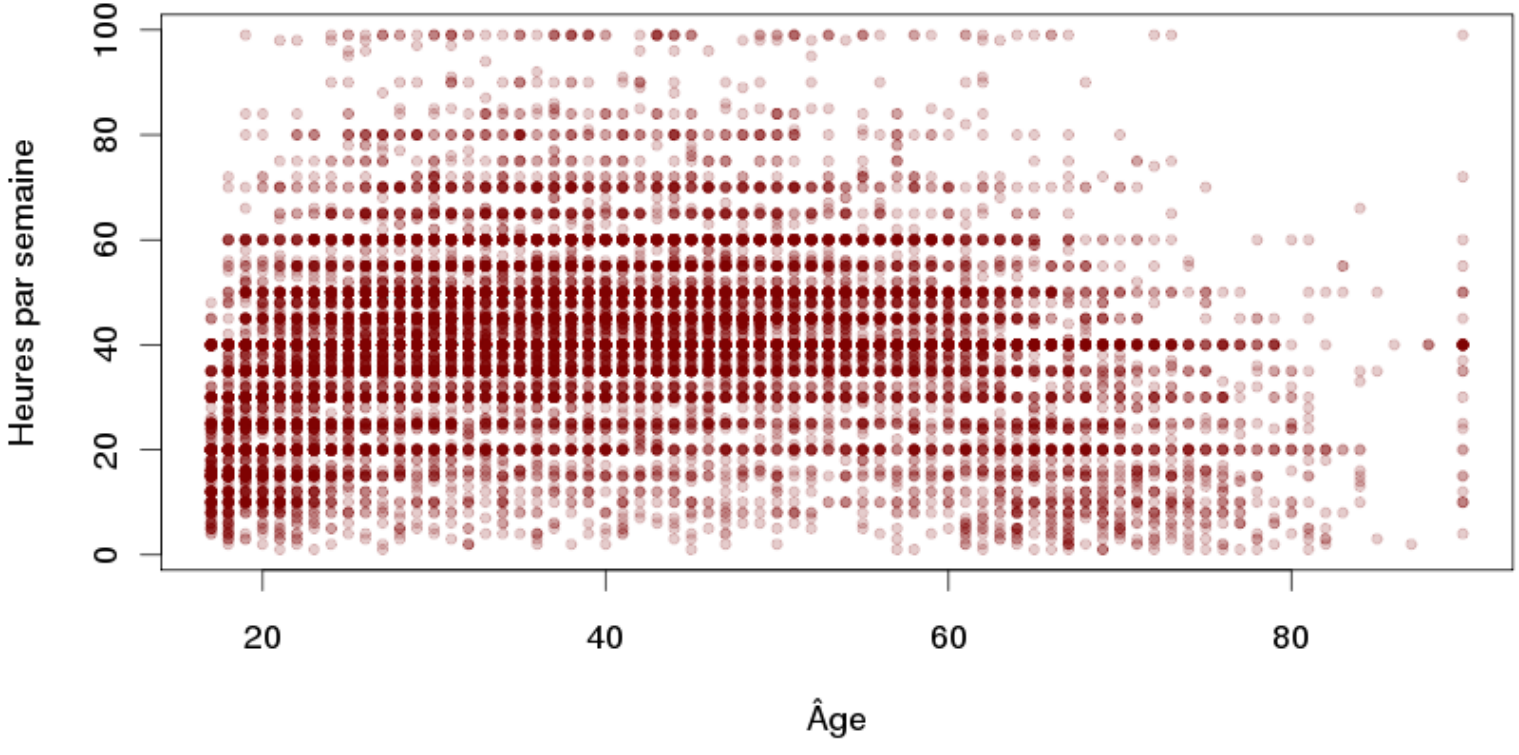
Plan

Analyses conjointes

- Relativement peu d'information dans chaque variable
- Analyse croisée nécessaire
- Difficultés :
  - vision humaine limitée (2D ou 3D, formes et couleurs)
  - beaucoup de combinaisons possibles
  - variables incompatibles
- Solutions :
  - outils de la visualisation de l'information (interaction)
  - outils de l'apprentissage automatique (automatisation)

Diagramme de dispersion

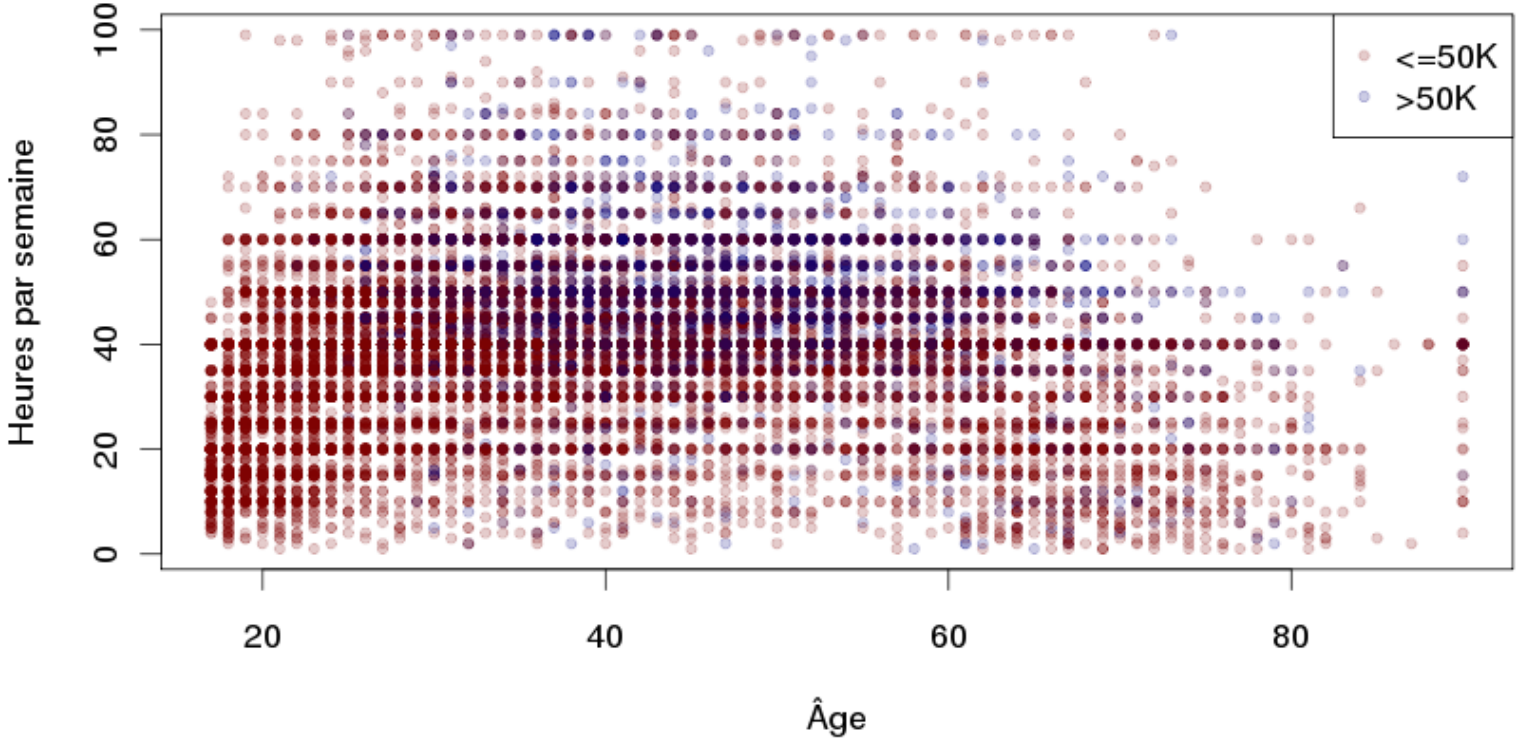
- Deux variables numériques : l'une en fonction de l'autre
- *scatter plot*



- Superposition : *alpha blending*

Décoration

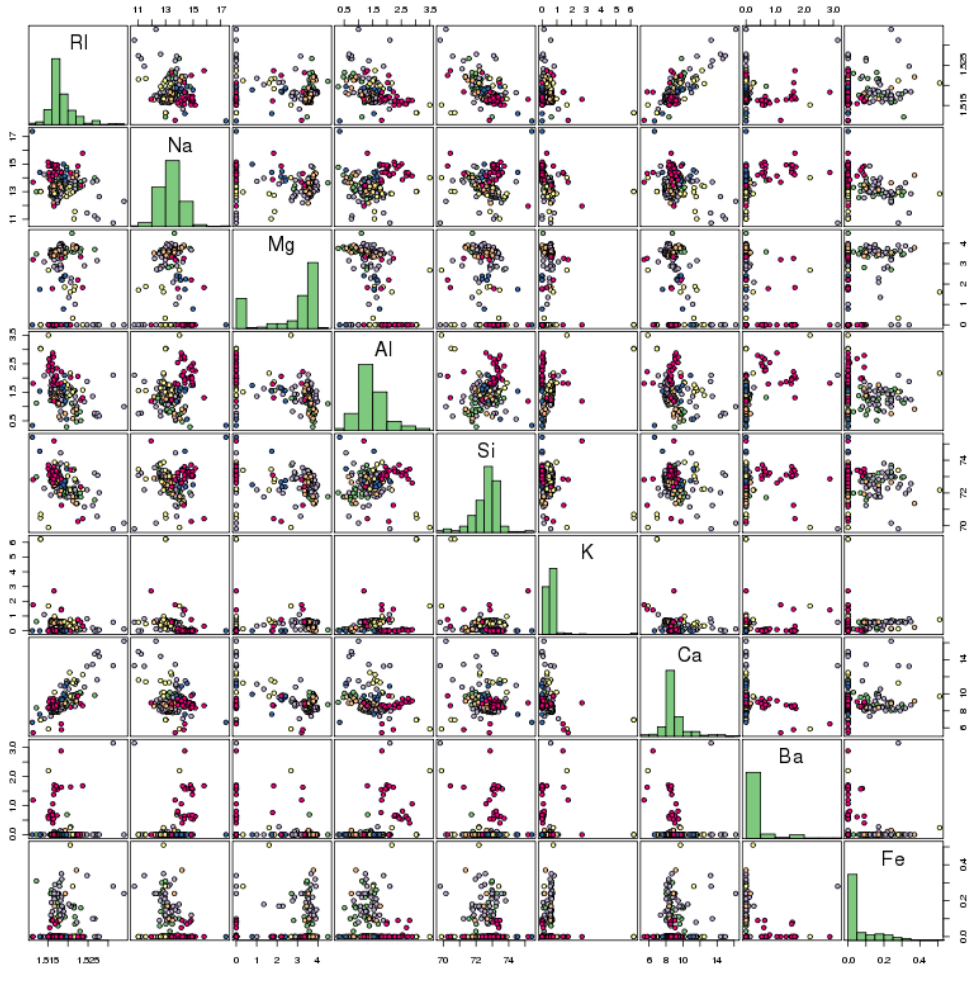
- Compléments du diagramme :
  - couleur en fonction d'une autre variable
  - symbole en fonction d'une autre variable



- Assez limité

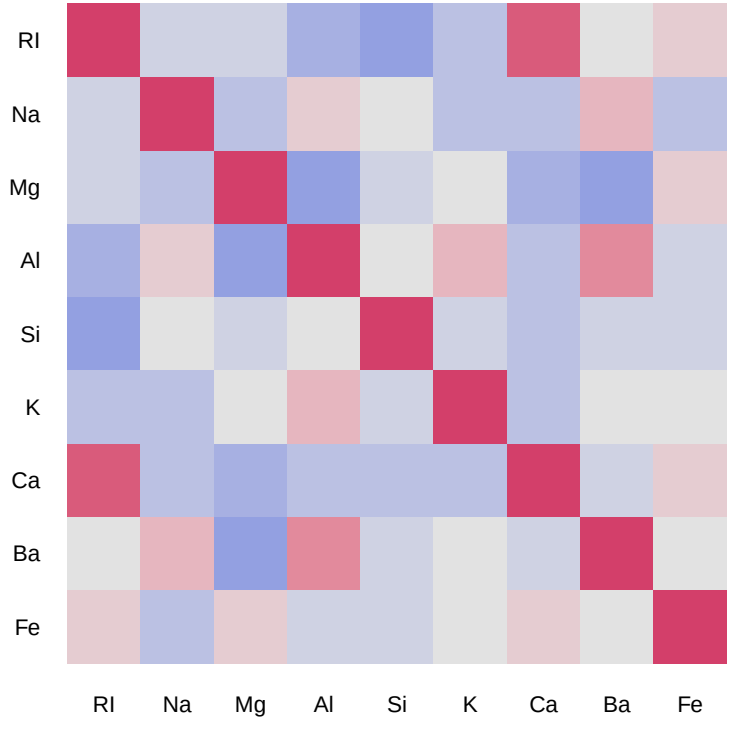
Matrice de diagrammes

- matrice de diagrammes de dispersion
- tous les couples de variables numériques
- limités à quelques variables (croissance quadratique)
- décorations possibles
- ici : 7 types de verre décrits par 9 variables



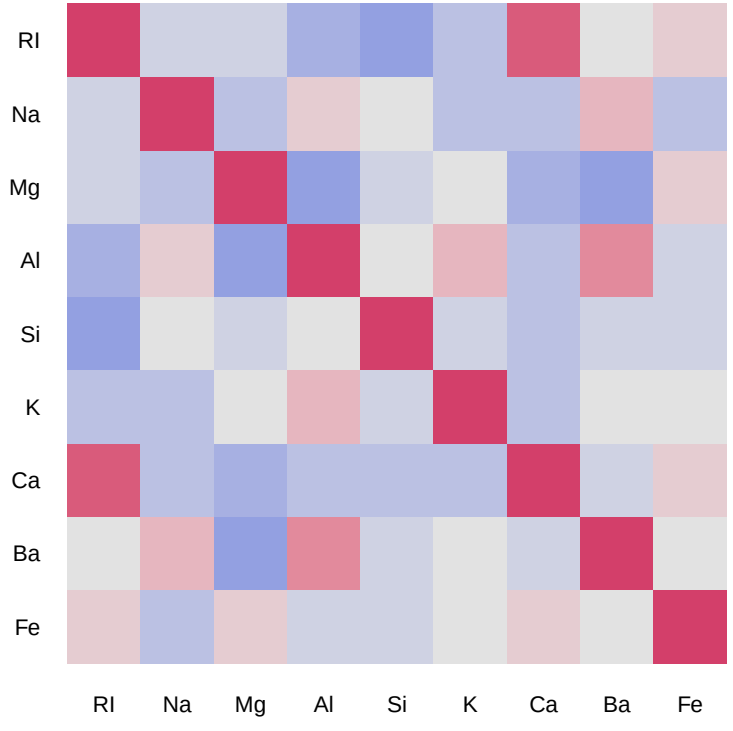
Corrélations

- Recherche de corrélations
- Représentation graphique de la matrice de corrélation :
  - rouge : forte corrélation positive
  - bleu : forte corrélation négative

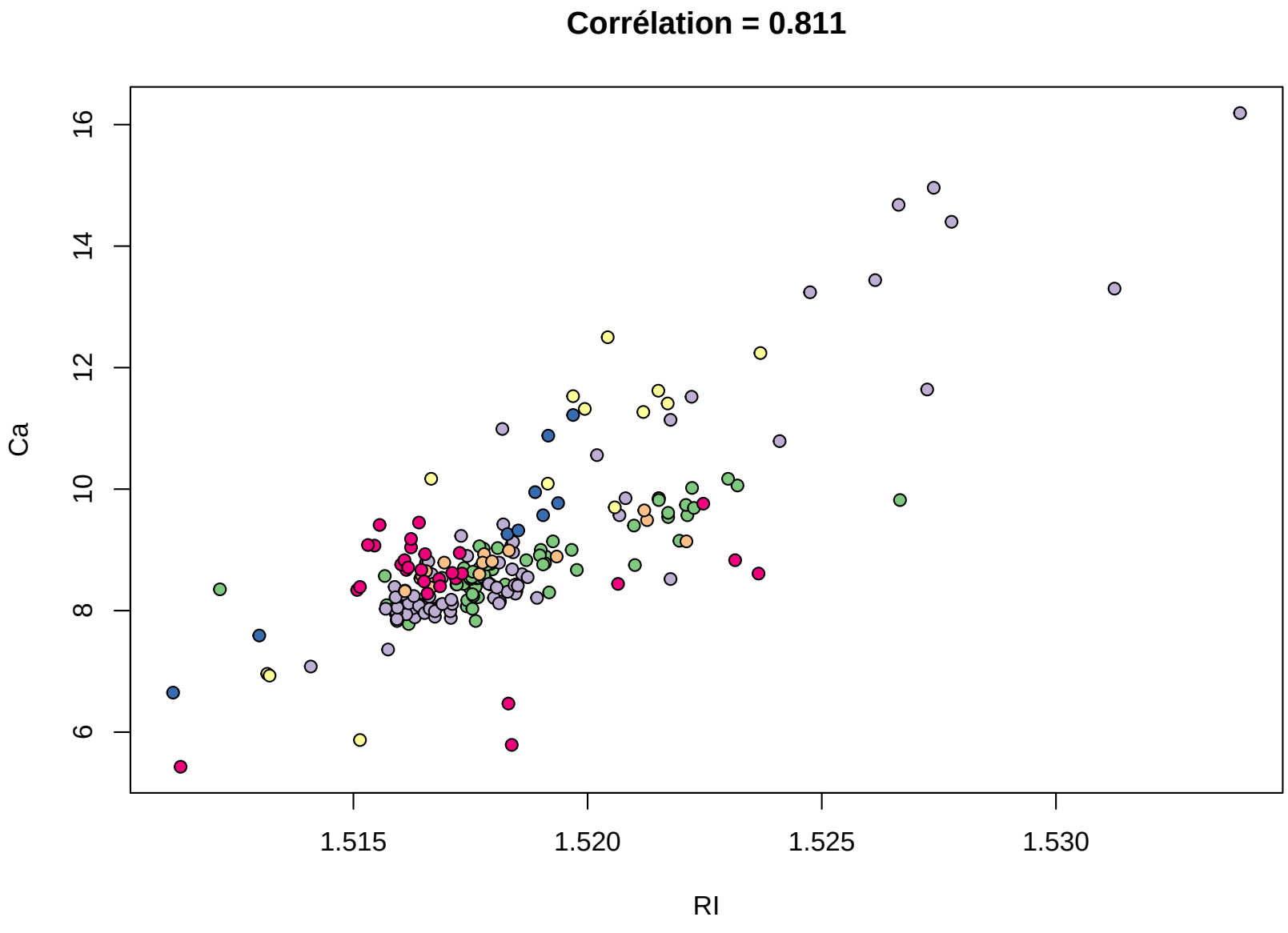


Corrélations

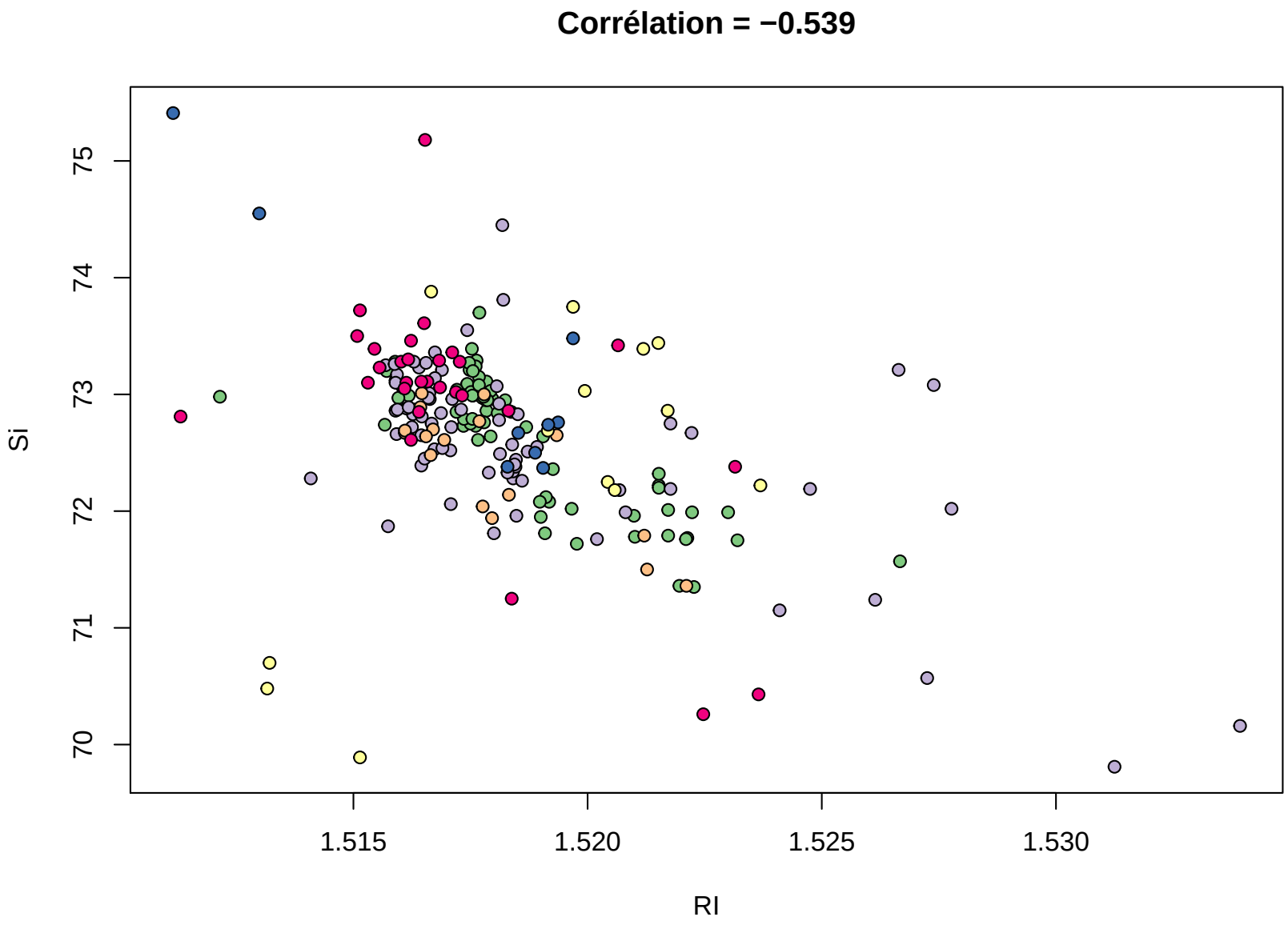
- Recherche de corrélations
- Représentation graphique de la matrice de corrélation :
  - rouge : forte corrélation positive
  - bleu : forte corrélation négative
- Ici :
  - RI corrélé avec Ca
  - Mg anti-corrélé avec Al
  - RI anti-corrélé avec Si
  - Aucun lien entre Al et Si



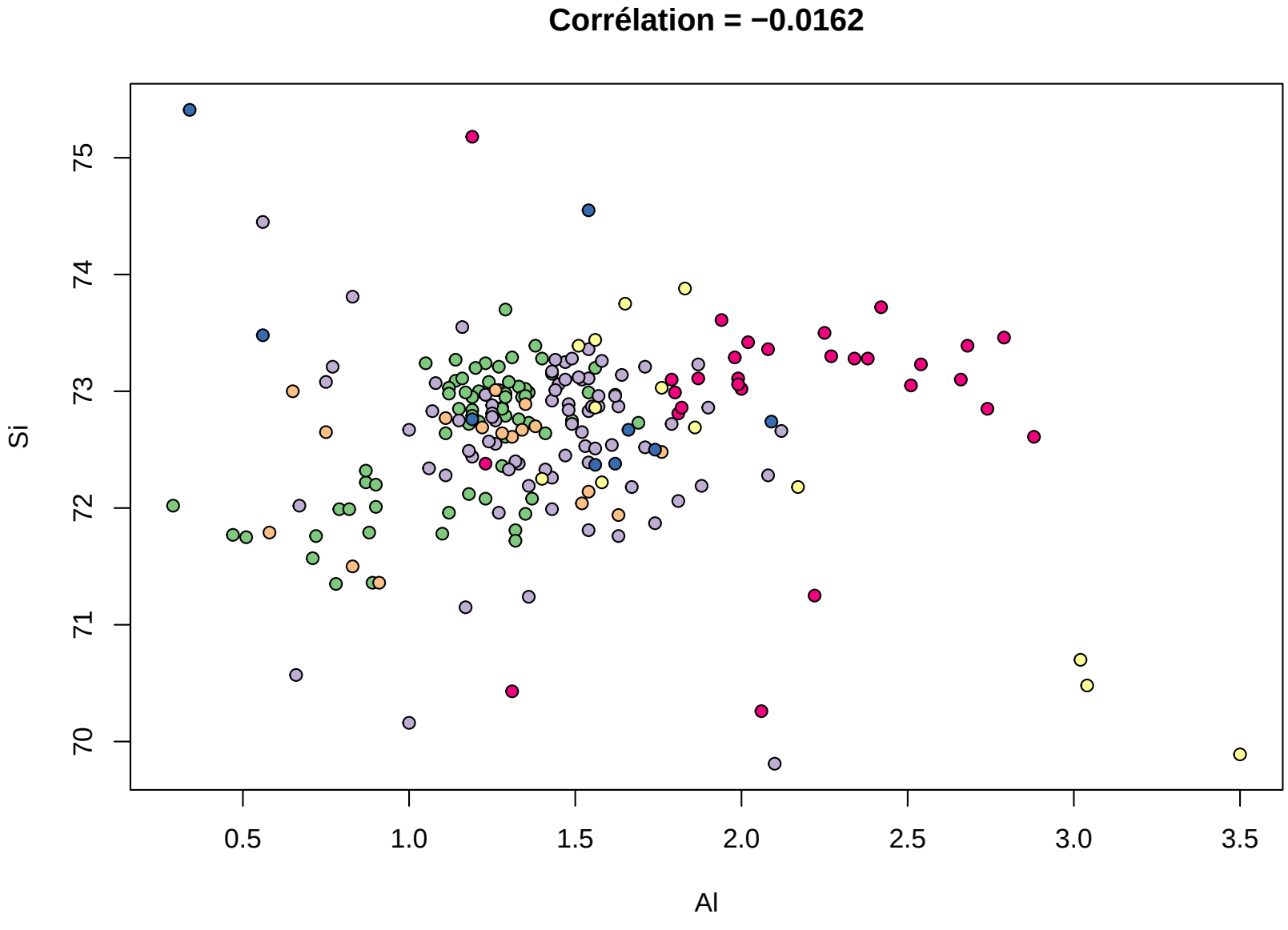
Corrélation RI et Ca



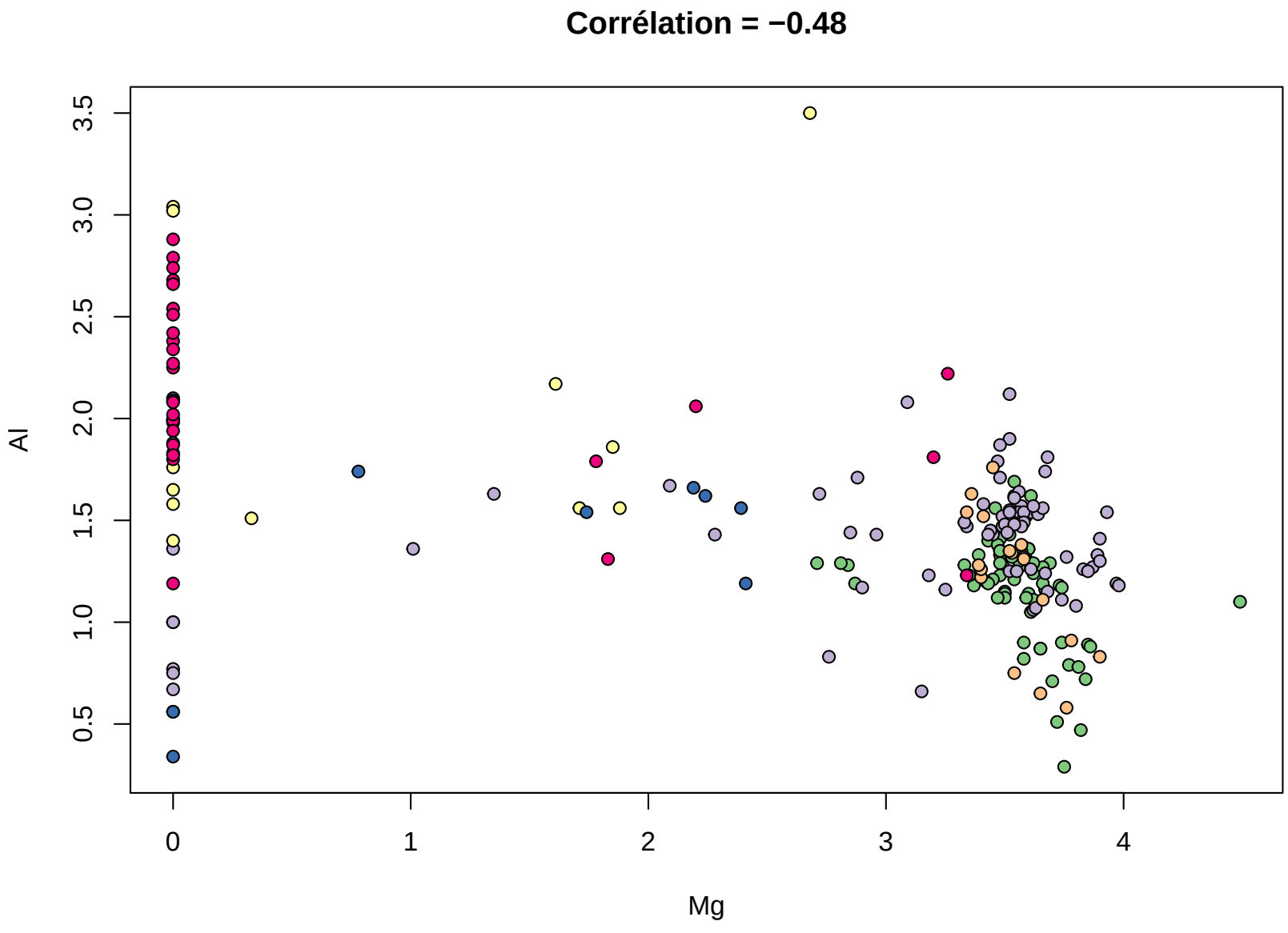
Corrélation RI et Si



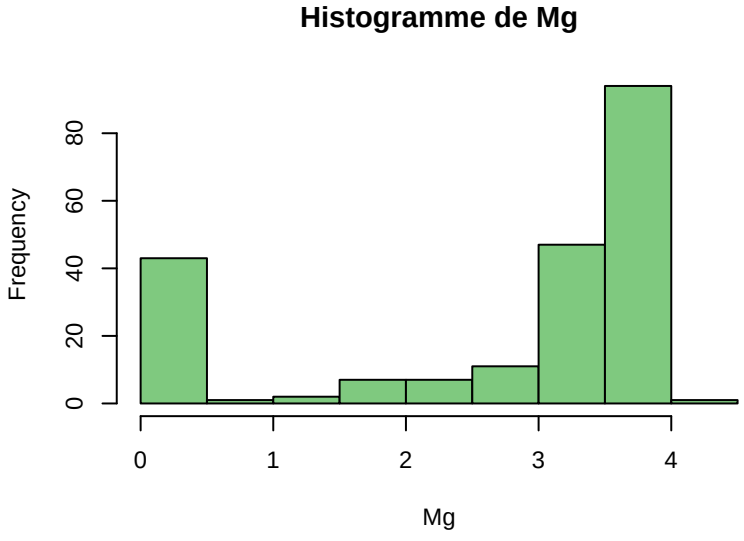
Corrélation Al et Si



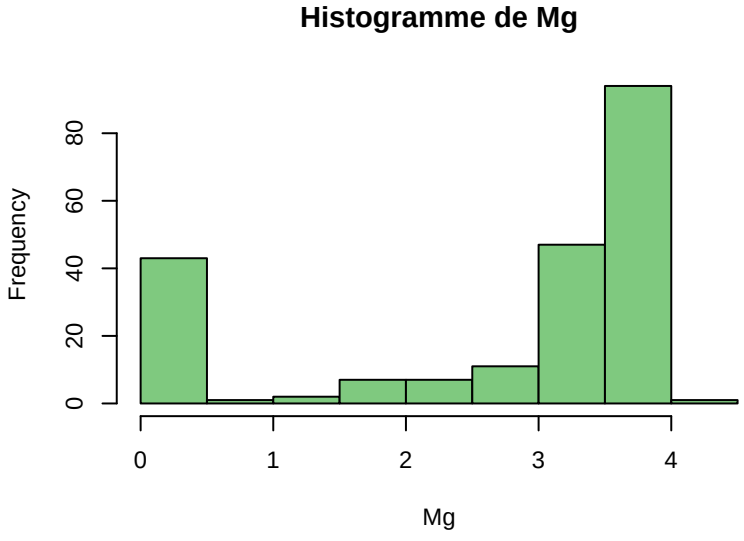
Corrélation Mg et Al

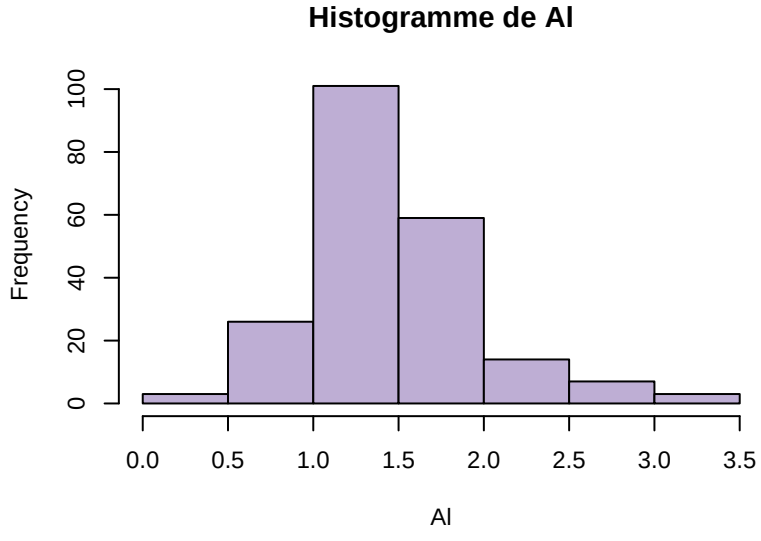


Mg et Al

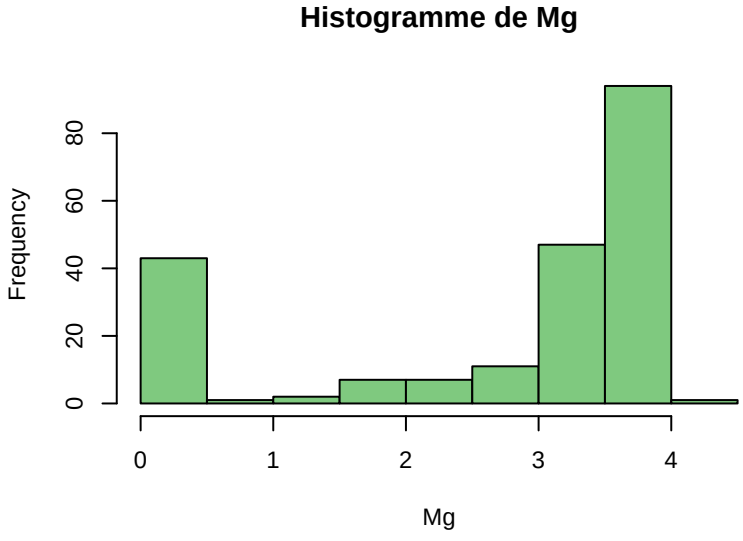


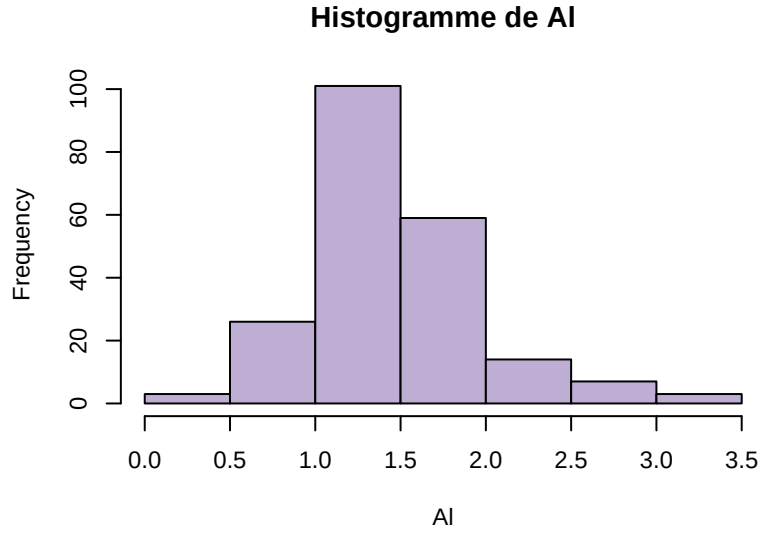
Mg et Al

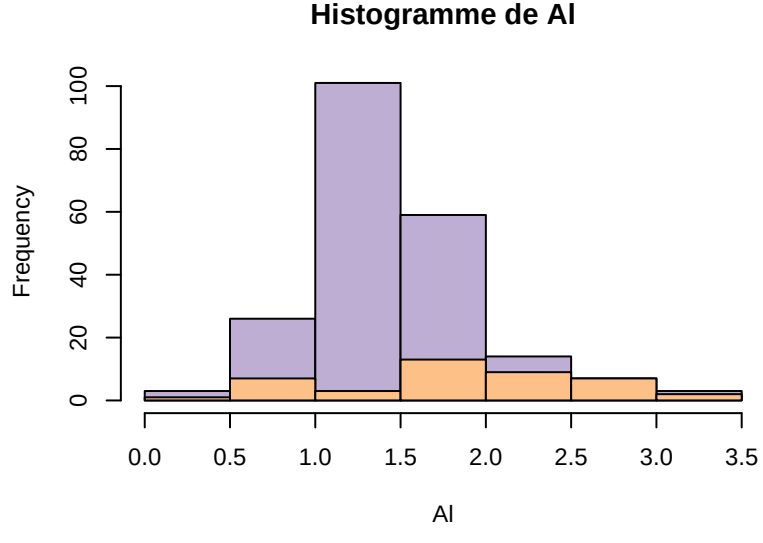

Histogramme de Mg
Frequency
0
20
40
60
80
0
1
2
3
4
Mg


Histogramme de Al
Frequency
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Al

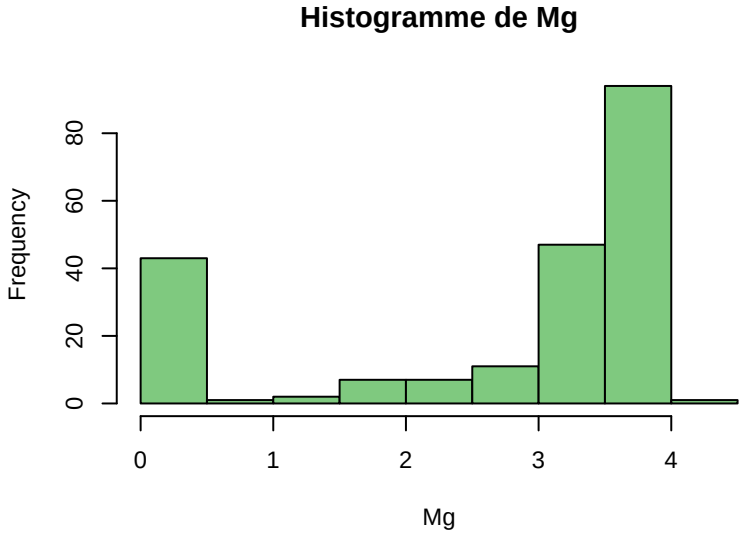
Mg et Al

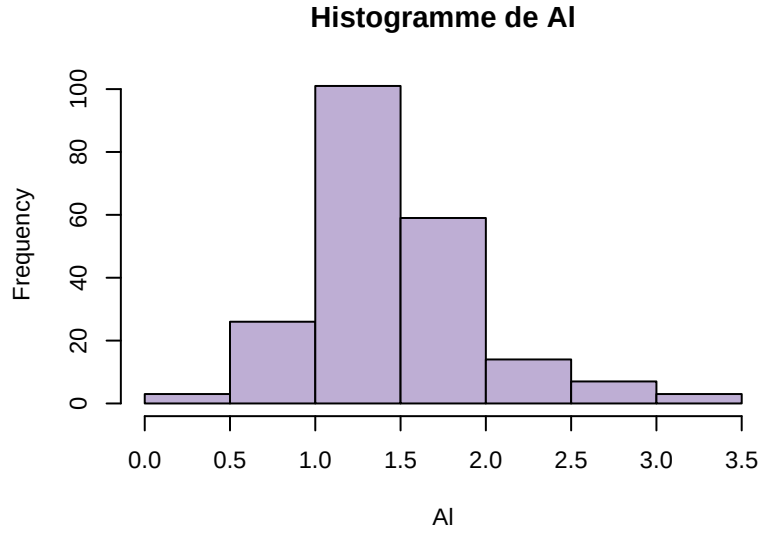

Histogramme de Mg
Frequency
0
20
40
60
80
0
1
2
3
4
Mg

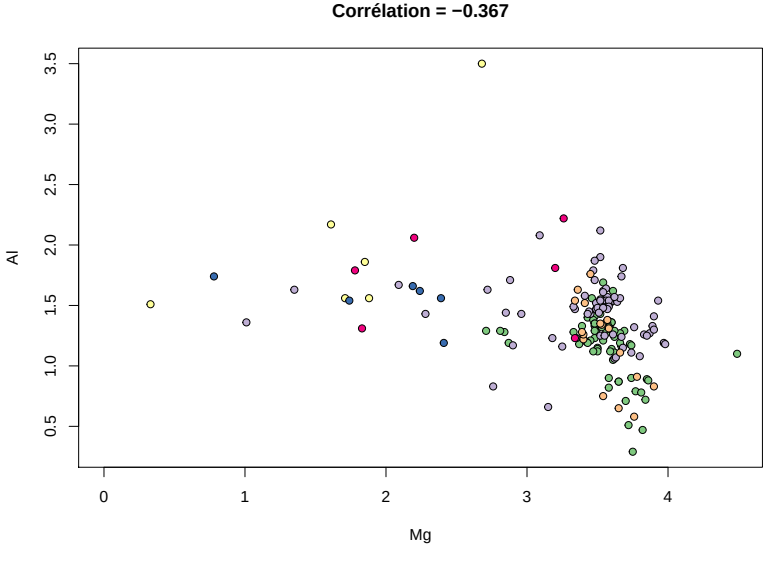

Histogramme de Al
Frequency
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Al

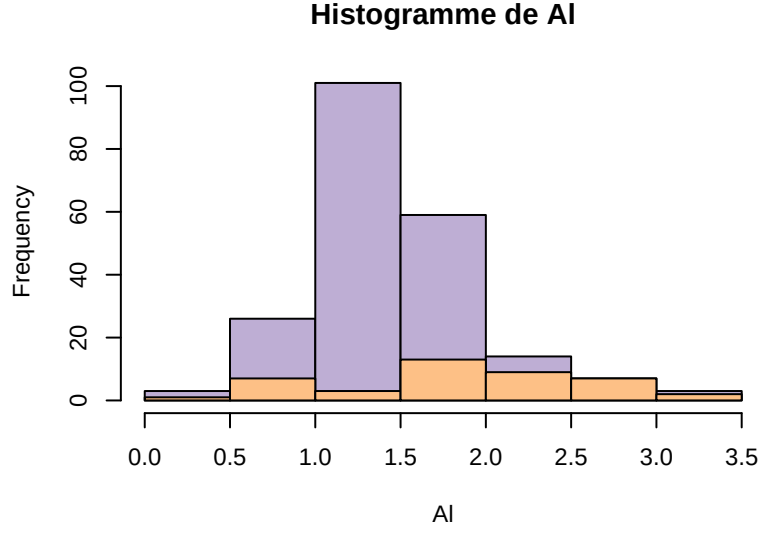

Histogramme de Al
Frequency
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Al

Mg et Al

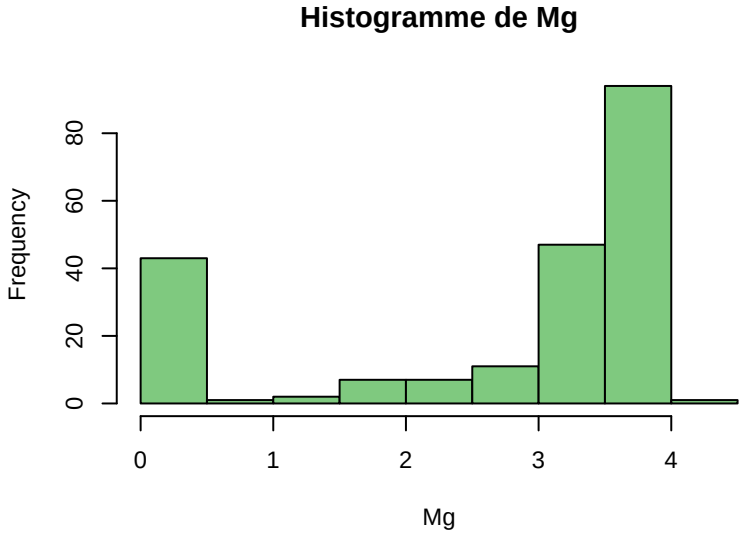

Histogramme de Mg
Frequency
0
20
40
60
80
0
1
2
3
4
Mg

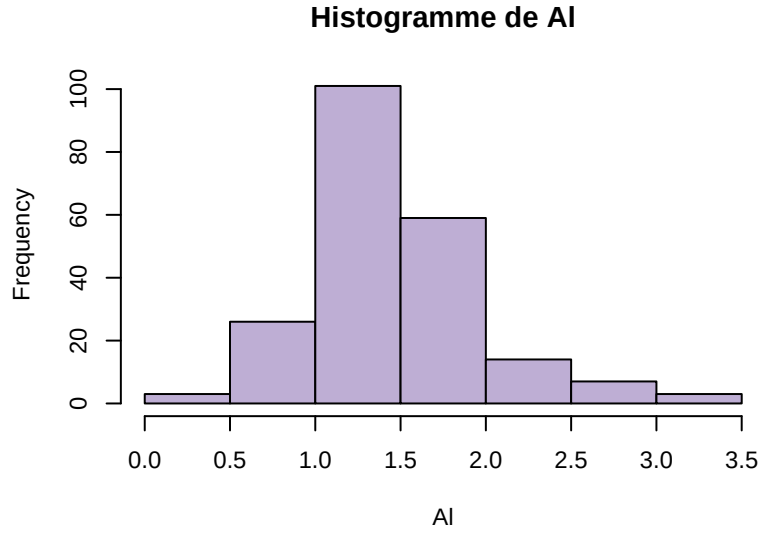

Histogramme de Al
Frequency
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Al

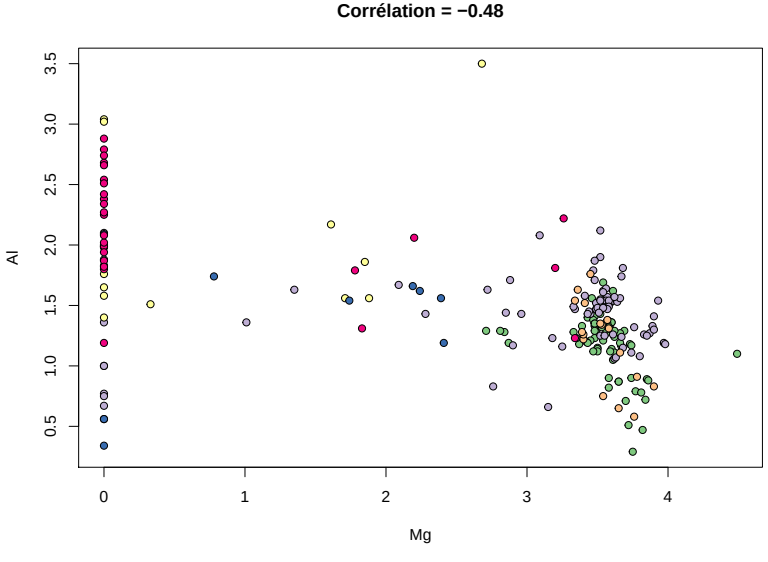

Corrélation = −0.367
Al
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
1
2
3
4
Mg

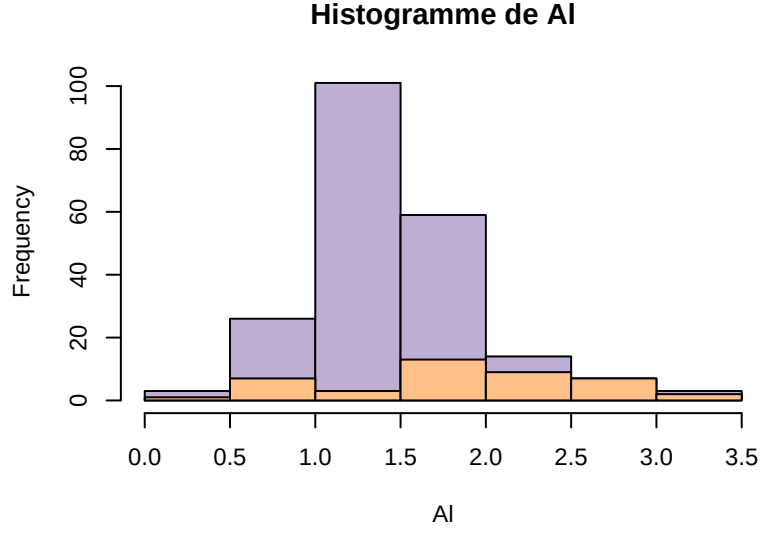

Histogramme de Al
Frequency
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Al

Mg et Al

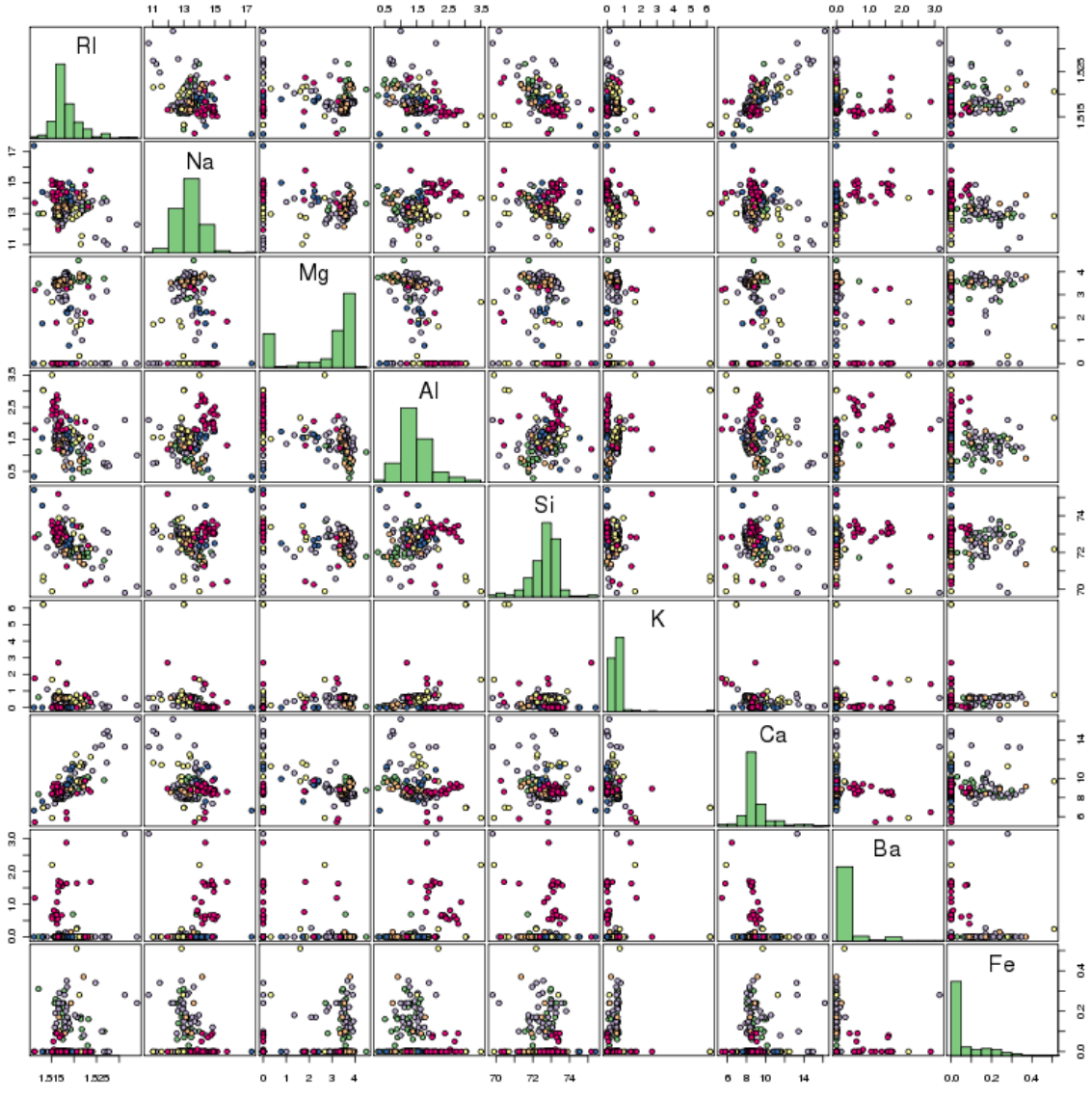

Histogramme de Mg
Frequency
0
20
40
60
80
0
1
2
3
4
Mg


Histogramme de Al
Frequency
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Al

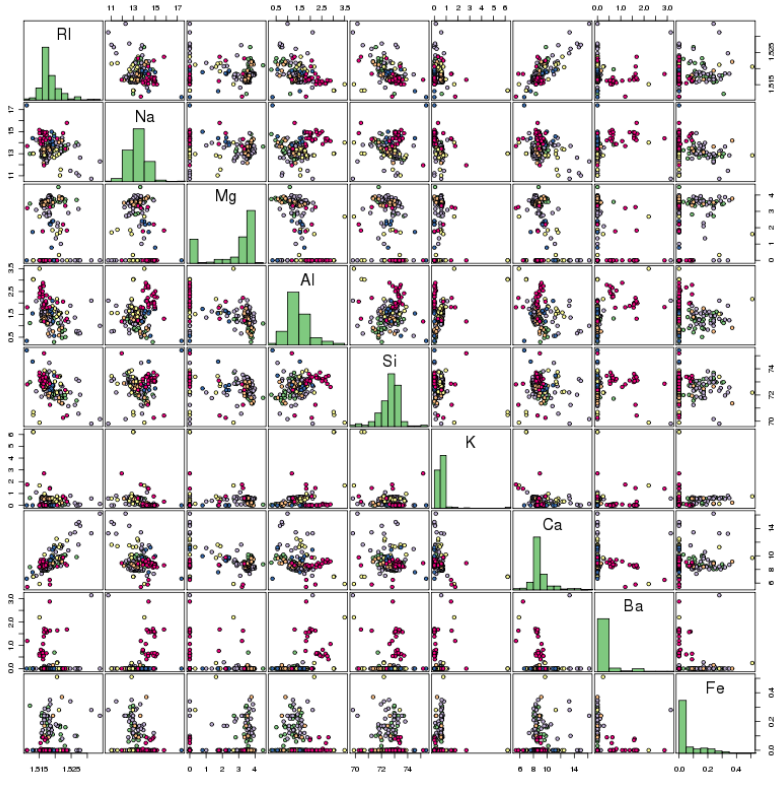

Corrélation = −0.48
Al
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0
1
2
3
4
Mg

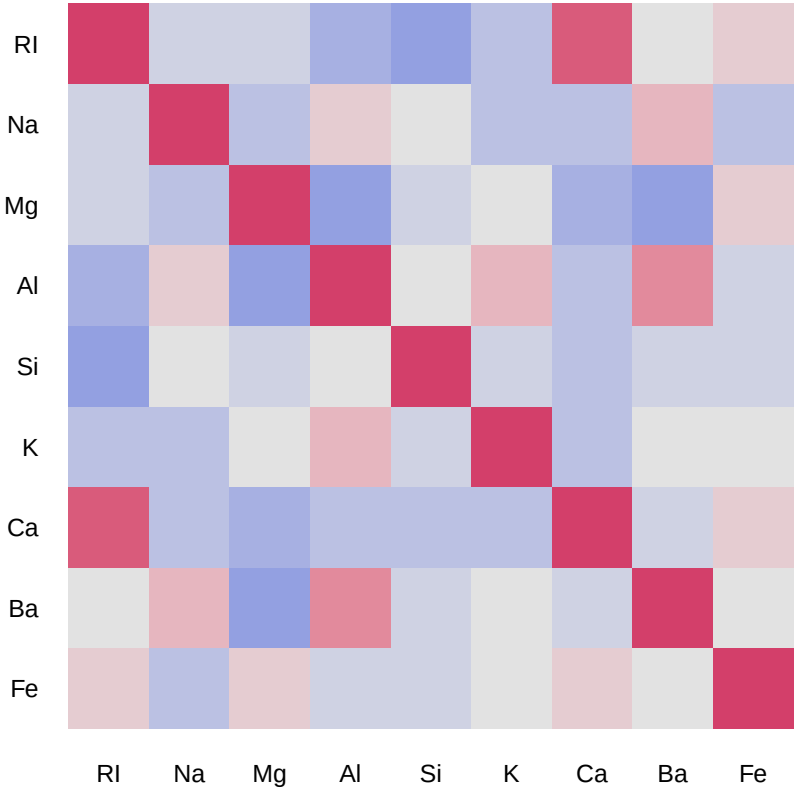

Histogramme de Al
Frequency
0
20
40
60
80
100
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Al

Vision globale



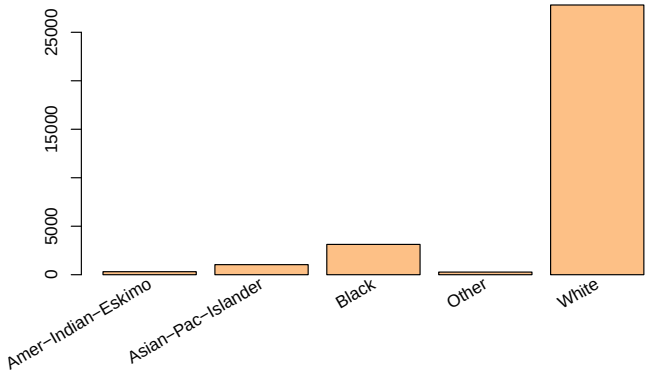
Vision globale



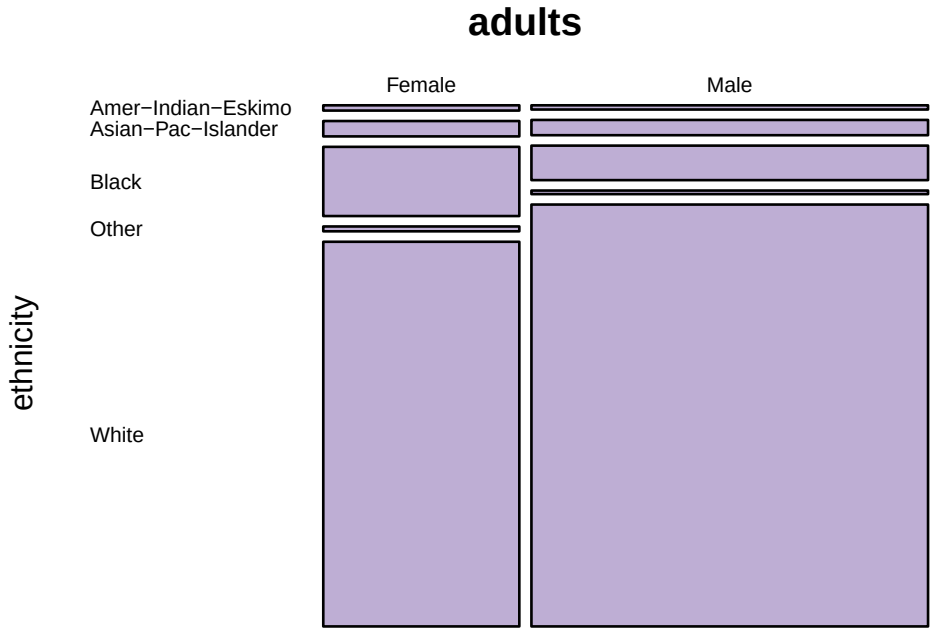


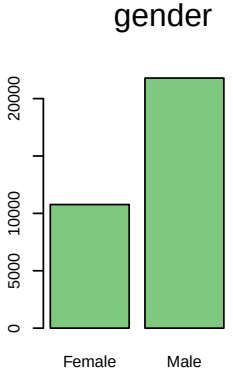
Mosaic plot

Équivalent du *scatter plot* pour les variables qualitatives



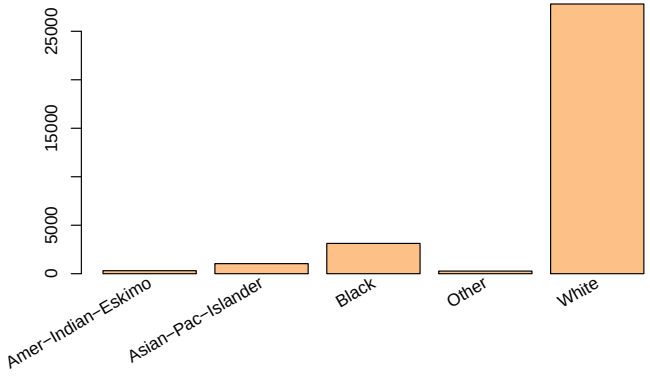
- découpage récursif
- surface proportionnelle à la fréquence



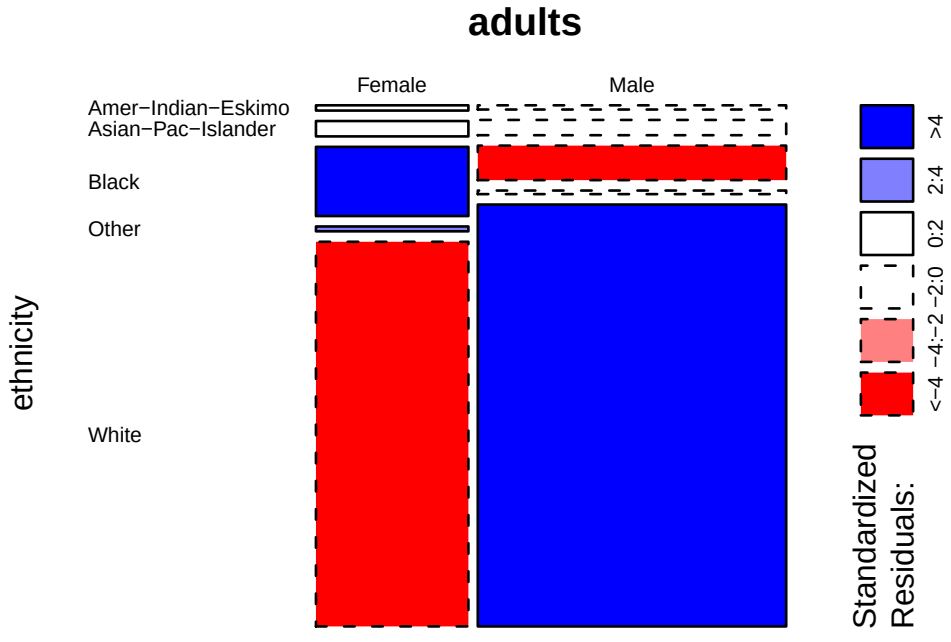


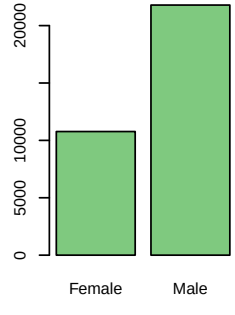
Mosaic plot

Équivalent du *scatter plot* pour les variables qualitatives



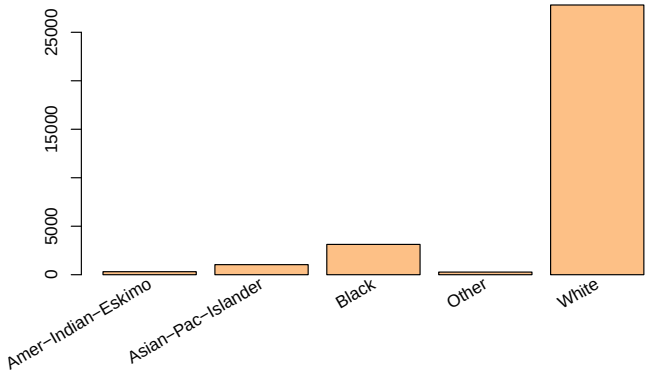
- découpage récursif
- surface proportionnelle à la fréquence
- significativité



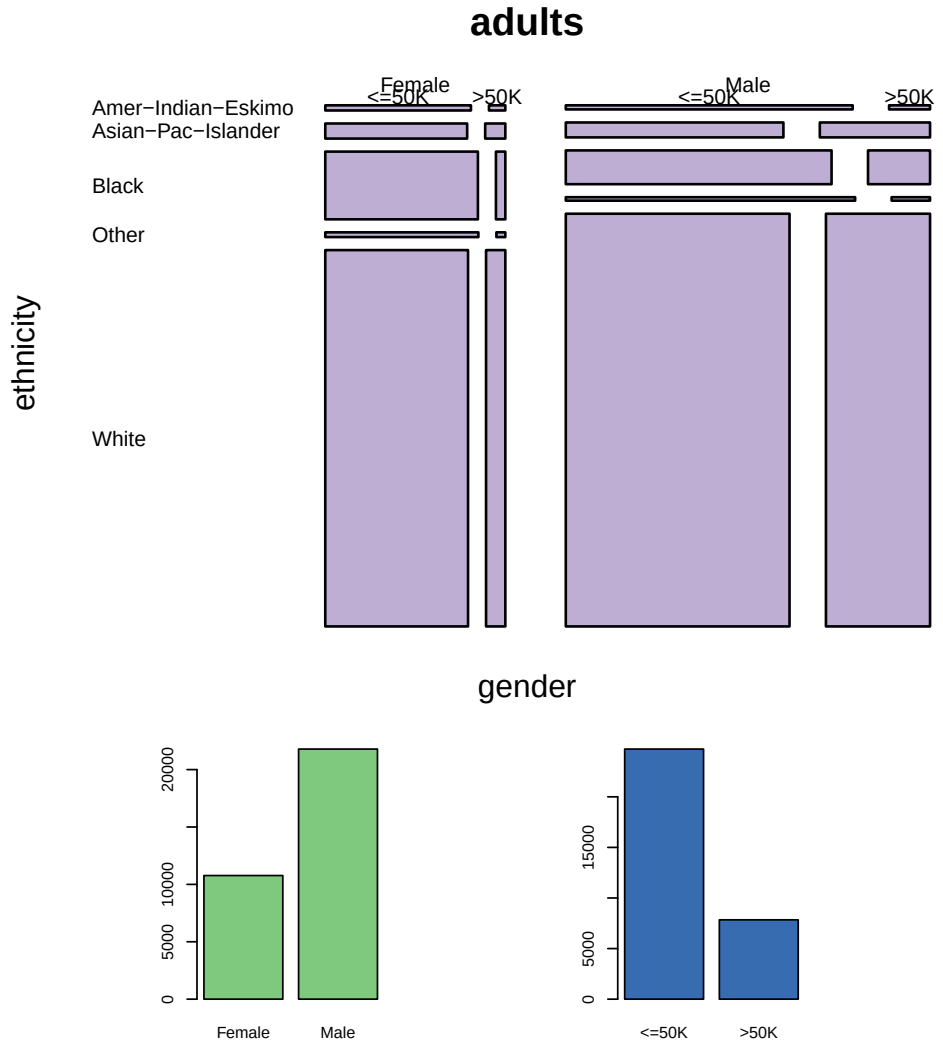


Mosaic plot

Équivalent du *scatter plot* pour les variables qualitatives

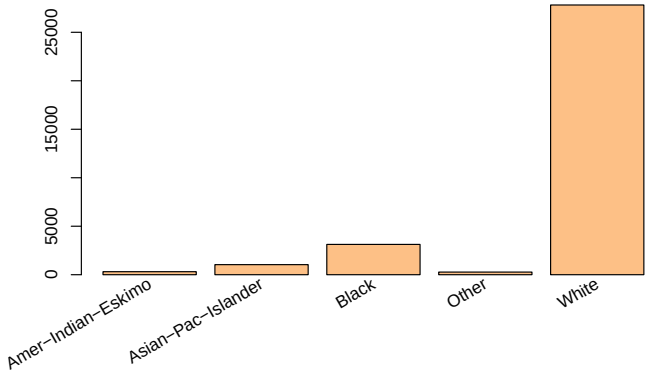


- découpage récursif
- surface proportionnelle à la fréquence
- significativité
- plus de 2 variables

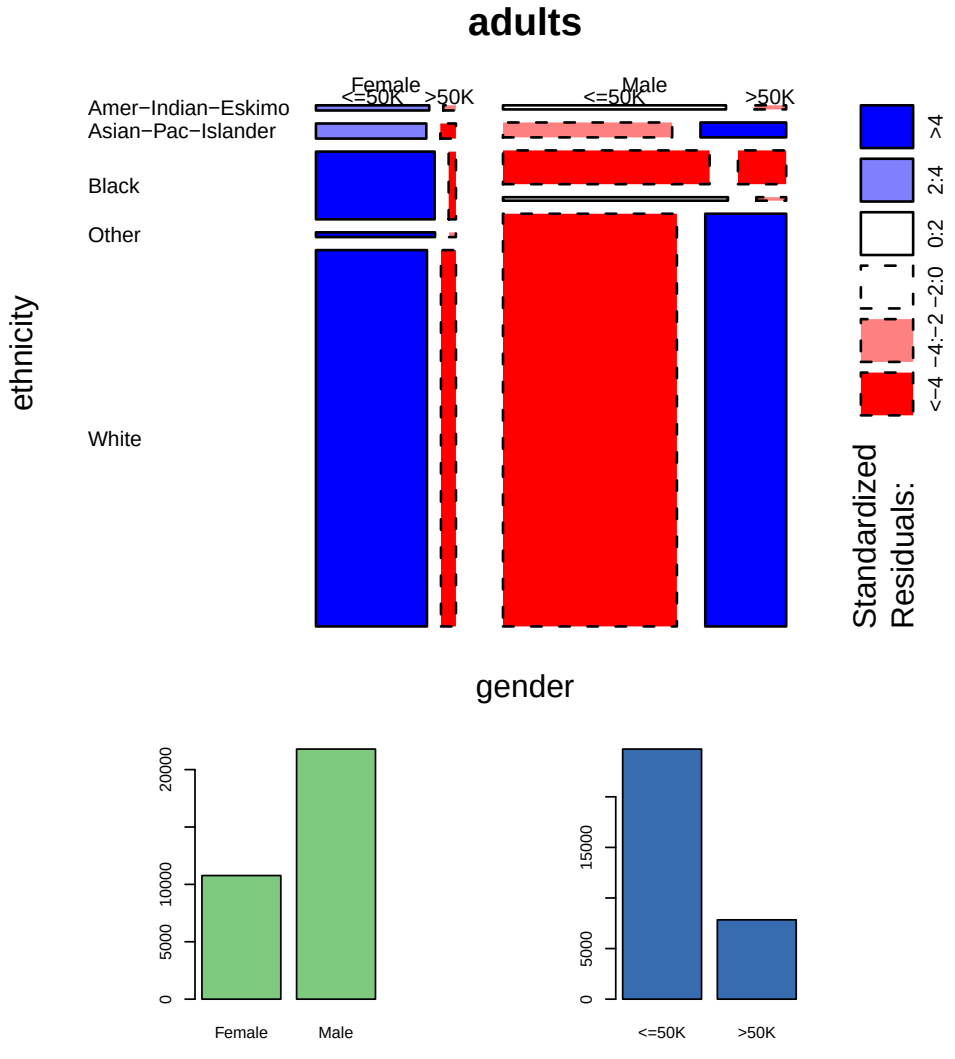


Mosaic plot

Équivalent du *scatter plot* pour les variables qualitatives



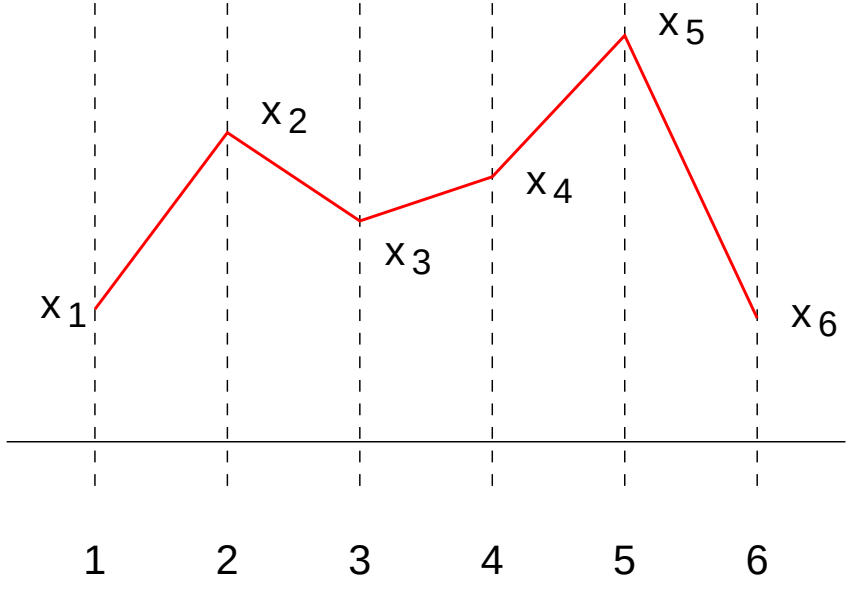
- découpage récursif
- surface proportionnelle à la fréquence
- significativité
- plus de 2 variables



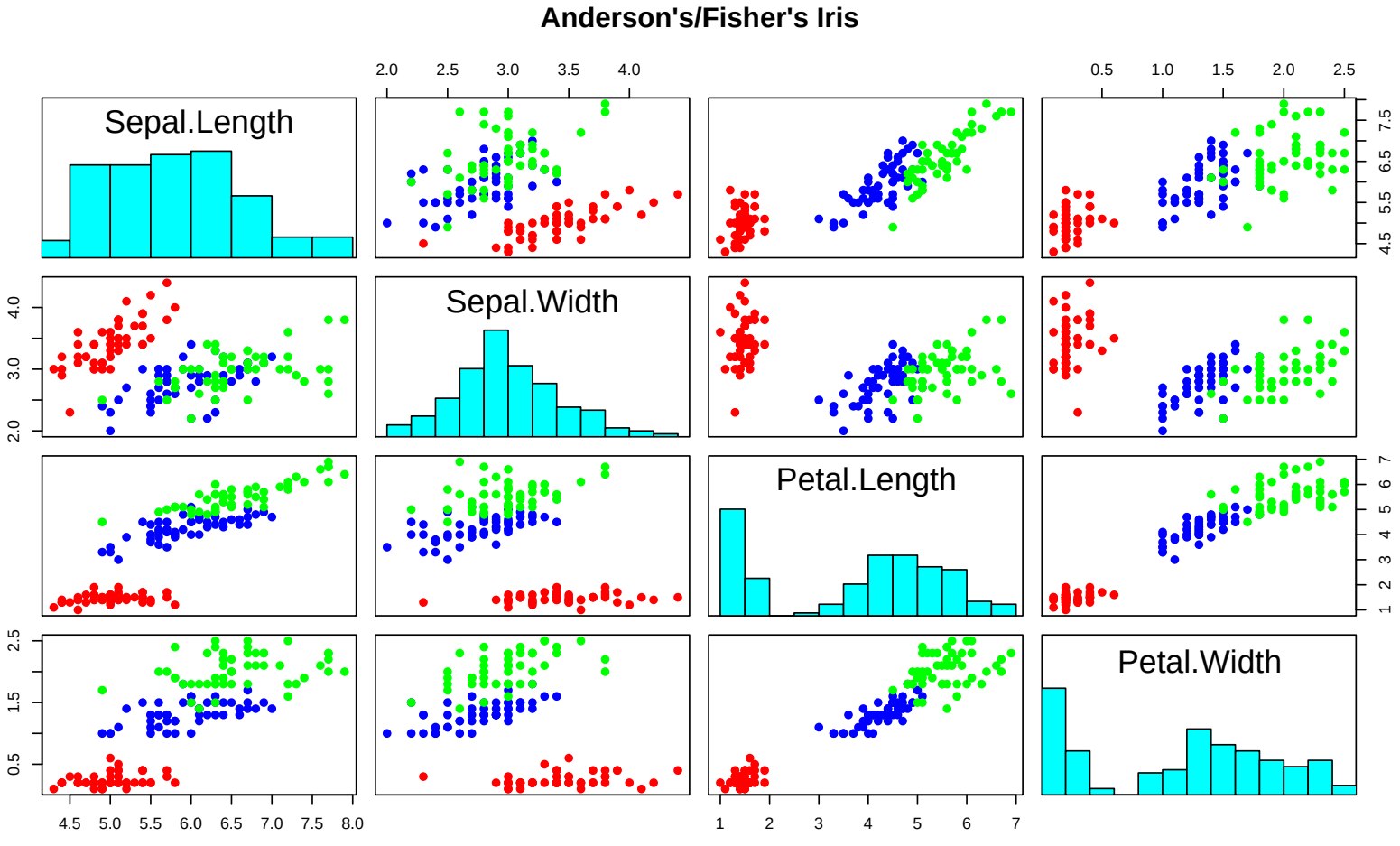
Coordonnées parallèles

Méthode proposée en 1985 par A. Inselberg

- un axe vertical par variable
- un objet devient une ligne brisée
- $(x_1, \ldots, x_p)$ est représenté par la ligne brisée passant par $(1, x_1)$, $(2, x_2)$, ..., (p, x_p)

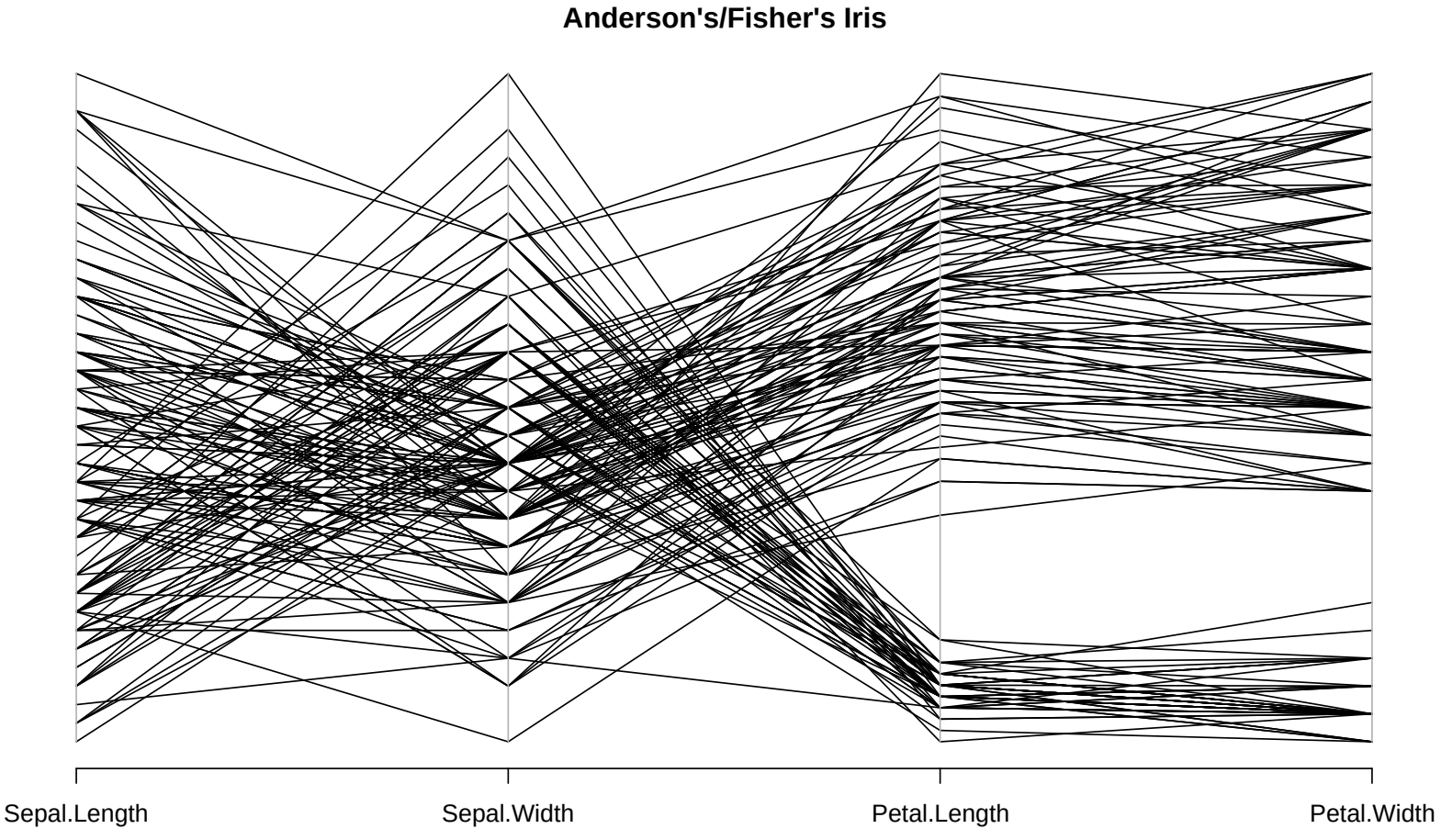


Données Iris

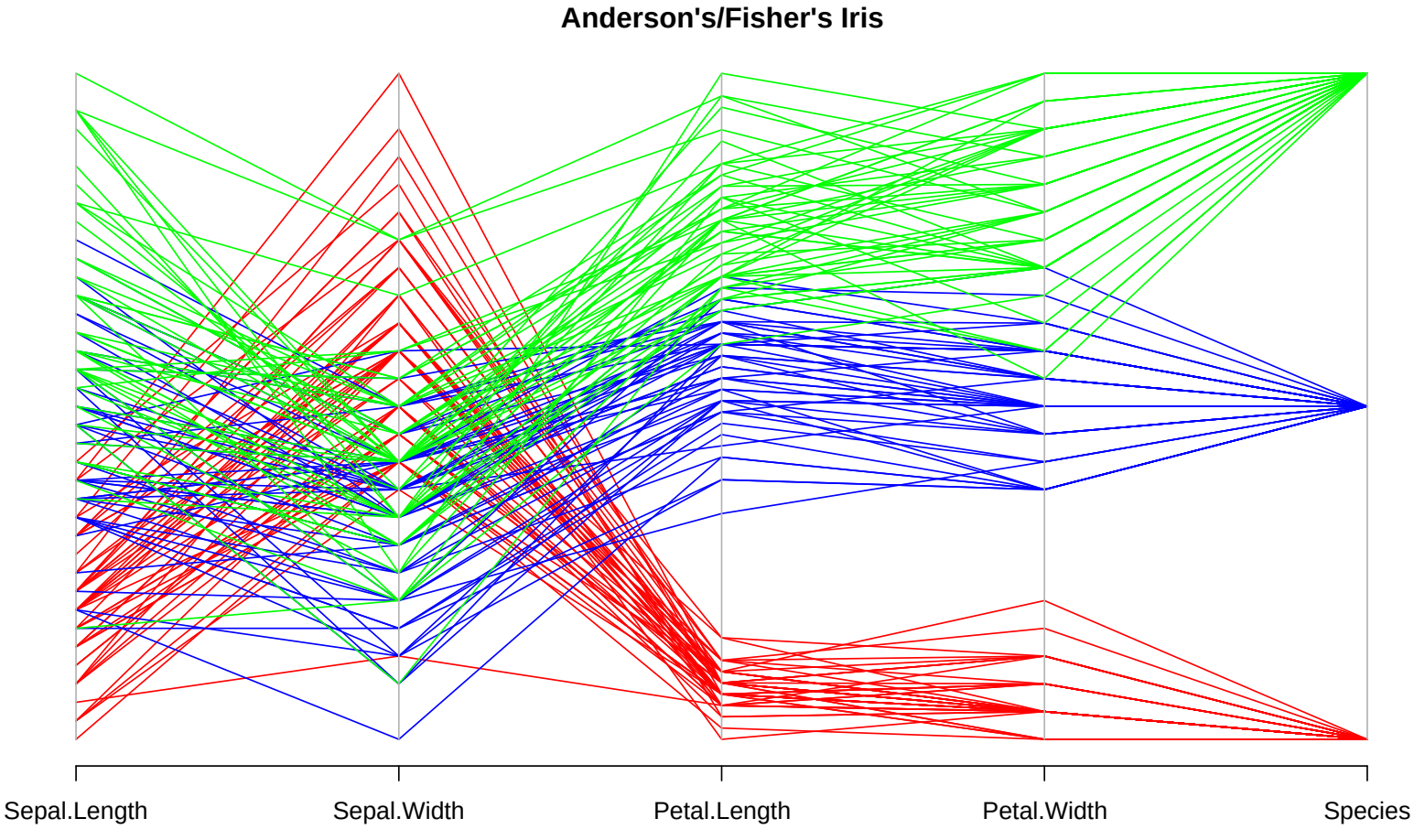


4+1 variables, 150 objets

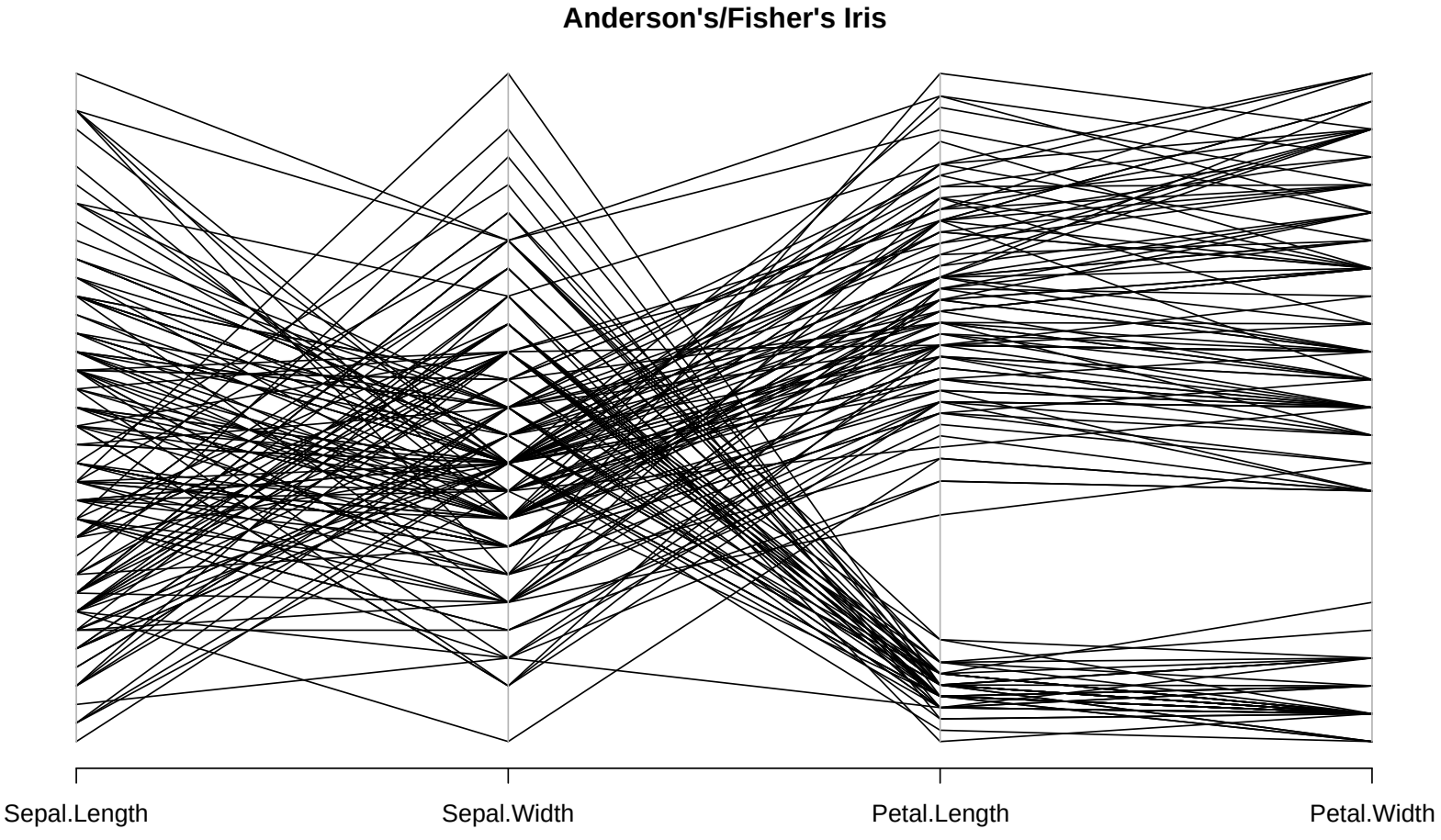
Données Iris



Données Iris



Attention à l'ordre



Les variables Petal sont elles corrélées ?

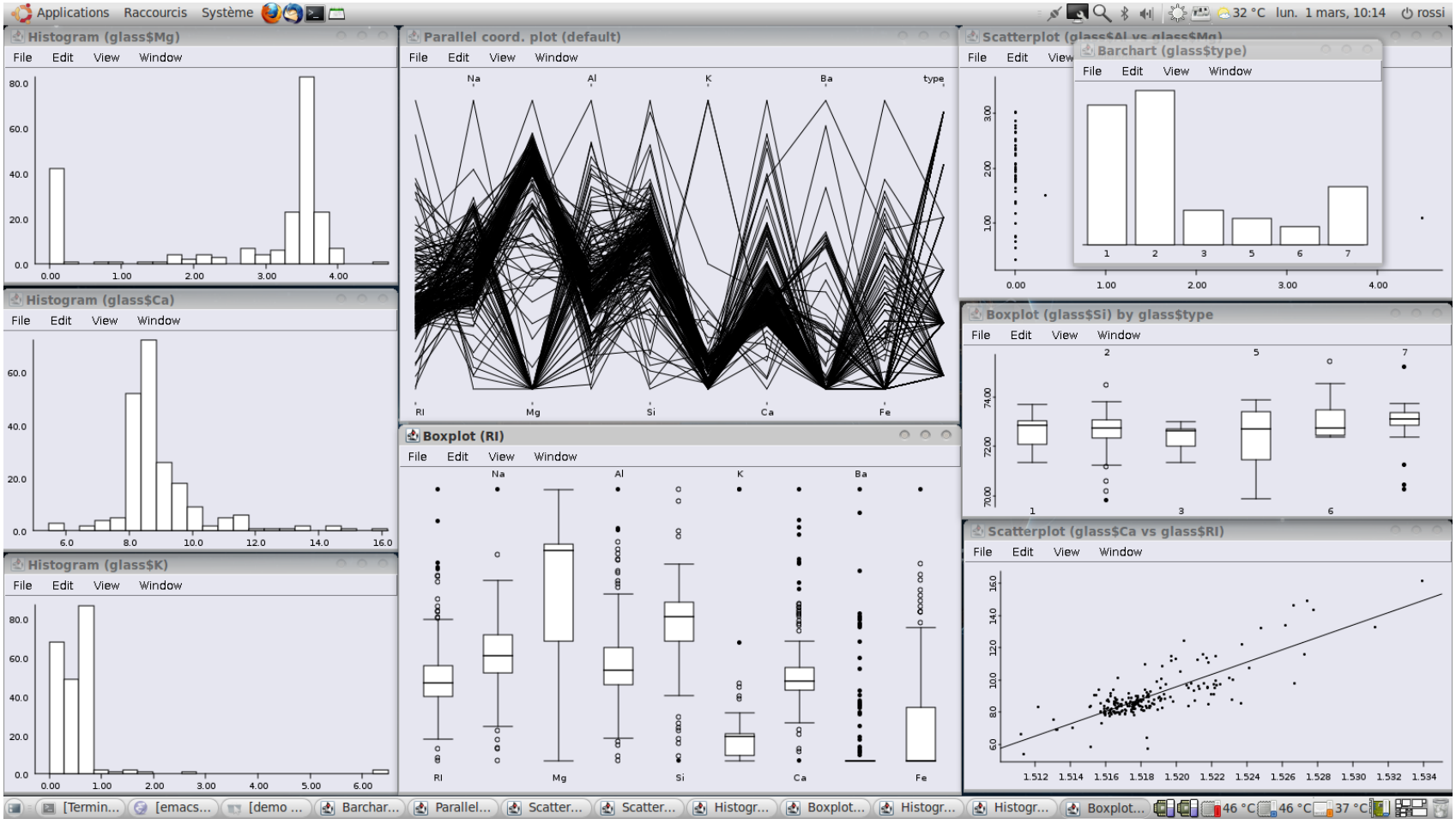
Attention à l'ordre

Les variables Petal sont elles corrélées ?

Interaction

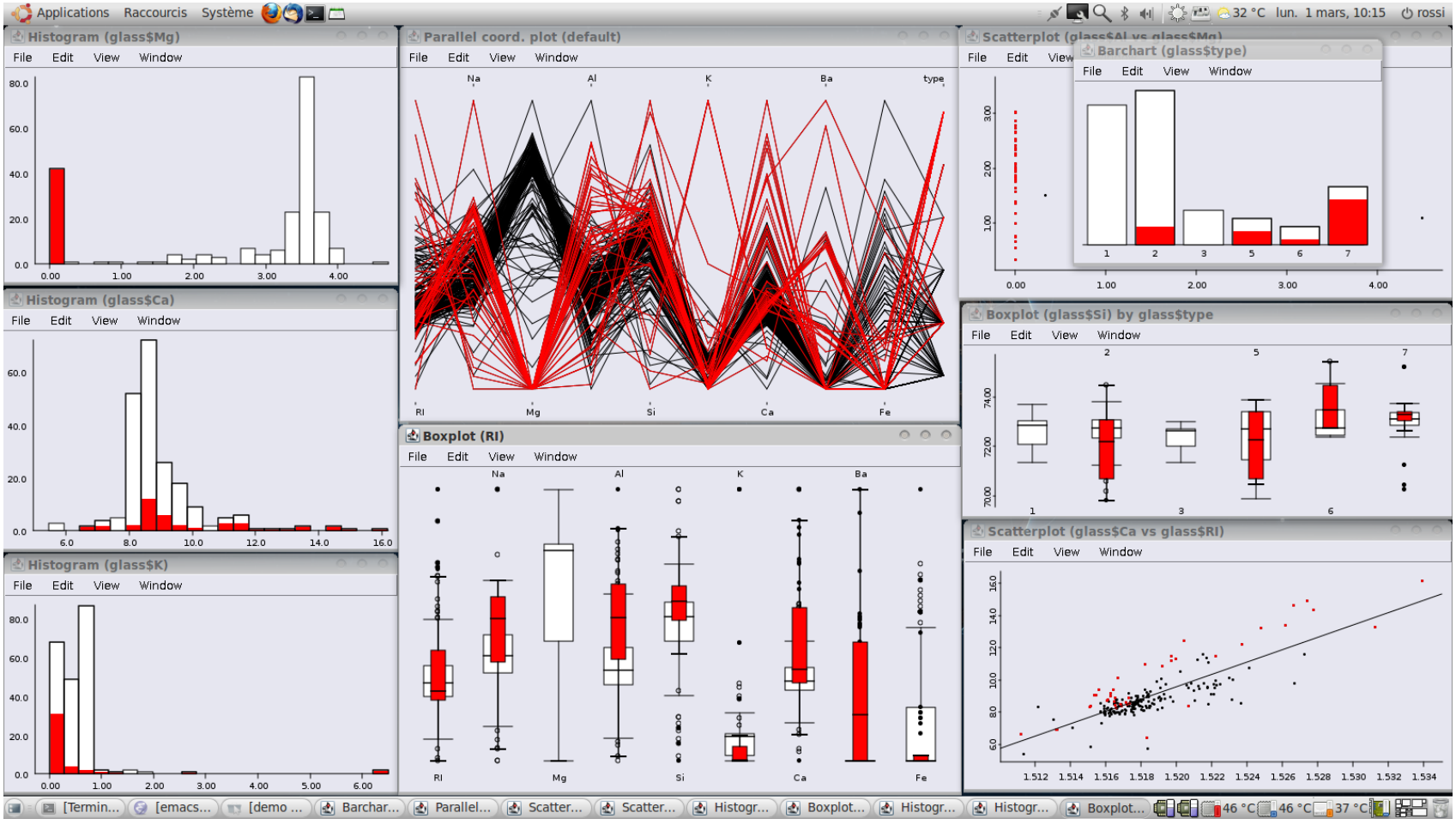
- problèmes :
  - surcharge de l'écran
  - surcharge cognitive
- solution par interaction :
  - zoom
  - vues multiples
  - sélection et lien :
    - sélection d'une zone (*brushing*)
    - affichage des résultats sur toutes les vues (*linking*)
- en R
  - iplots
  - ggobi et rggobi

iplots



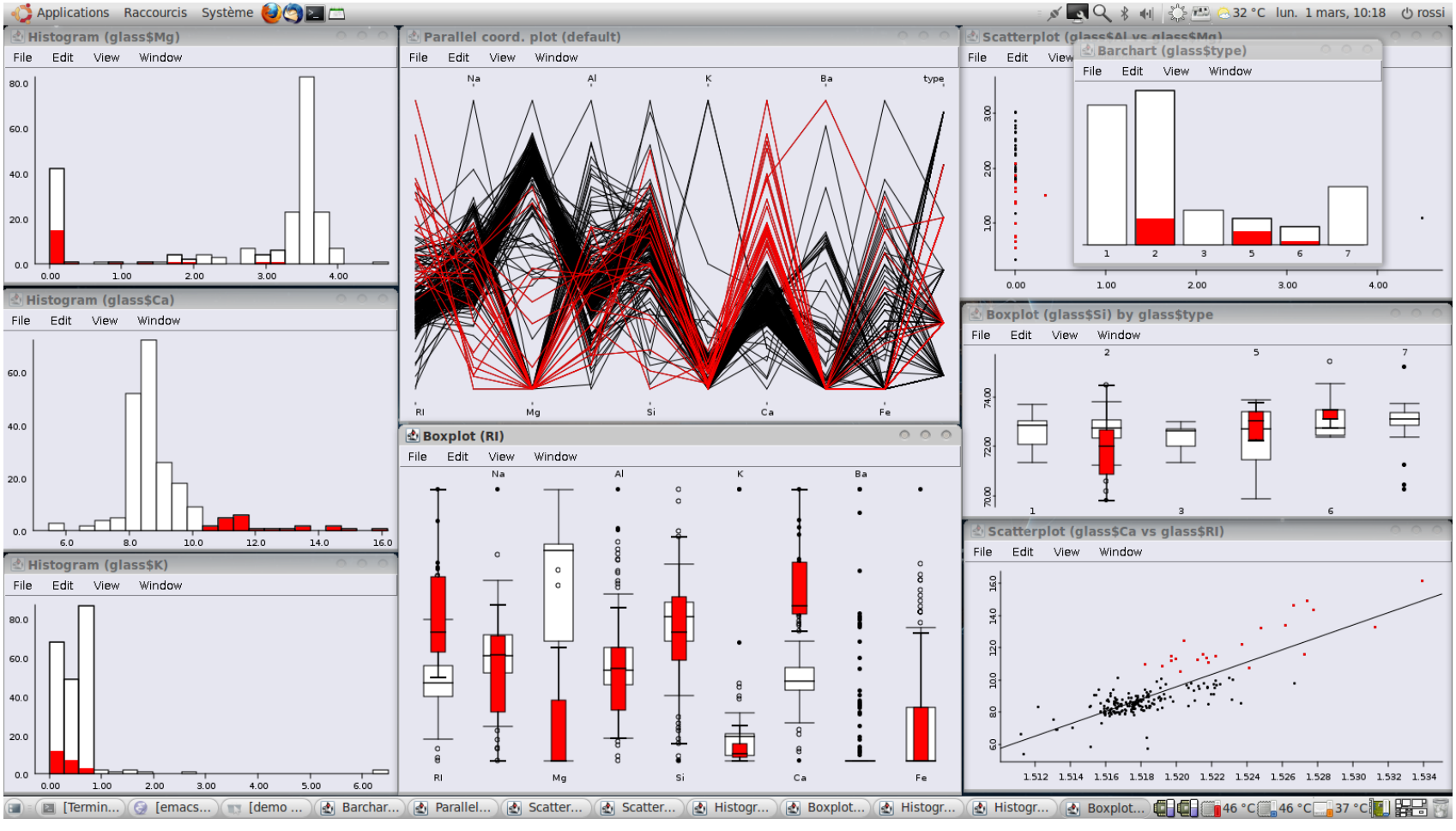
TELECOM ParisTech

iplots

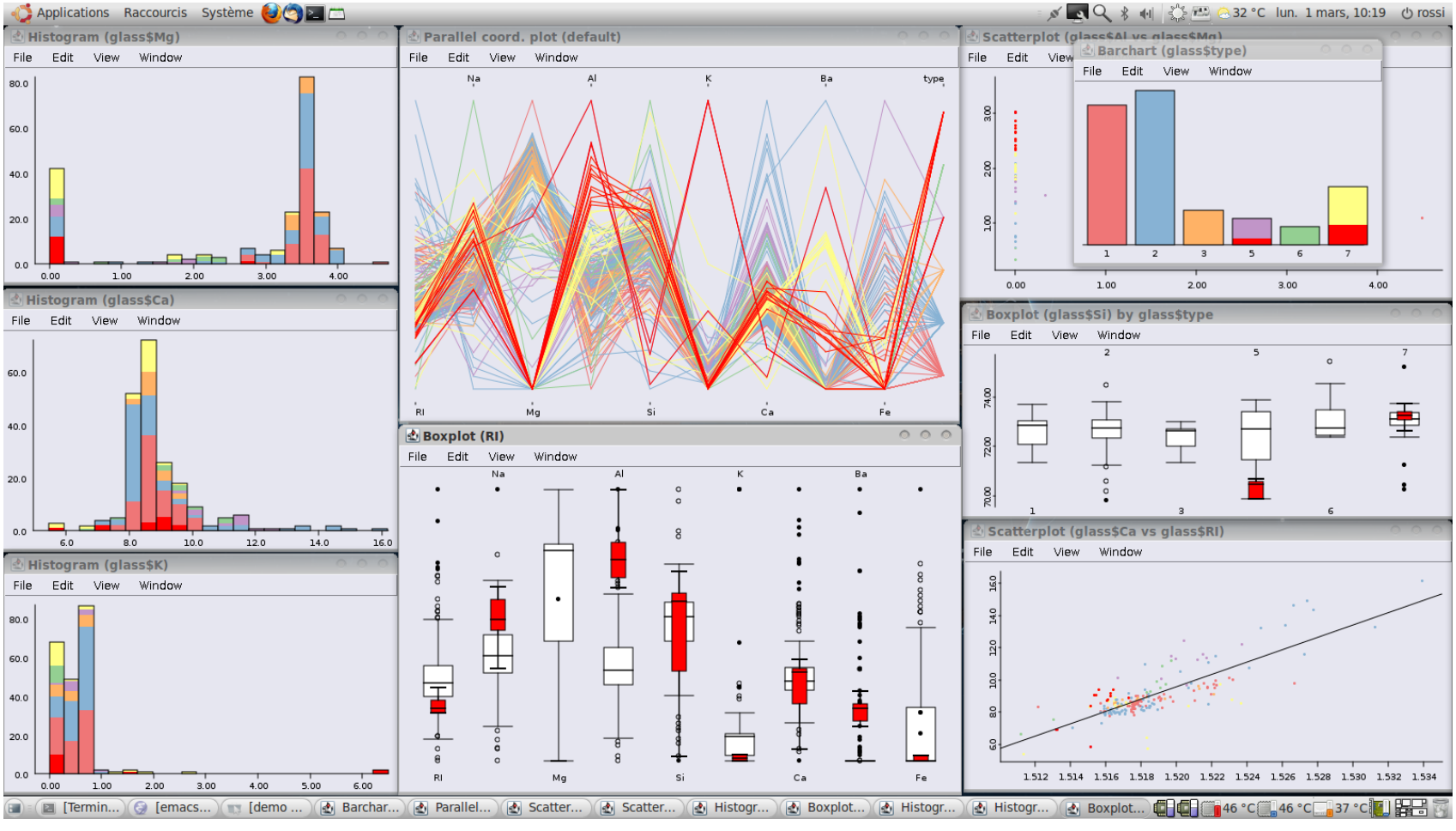


TELECOM ParisTech

iplots


Applications Raccourcis Système
32 °C lun. 1 mars, 10:18 rossi
Histogram (glass$Mg)
Parallel coord. plot (default)
Scatterplot (glass$Al vs glass$Mg)
Barchart (glass$type)
Histogram (glass$Ca)
Boxplot (glass$Si) by glass$type
Boxplot (RI)
Histogram (glass$K)
Scatterplot (glass$Ca vs glass$RI)
File Edit View Window
Na Al K Ba type
RI Mg Si Ca Fe

iplots



TELECOM ParisTech