Royaume du Maroc

Ministère de l'Habitat, de l'Urbanisme et de la Politique de la ville

Partie II : Les statistiques avec R

1. Objectifs

L'objet de cette formation est de vous permettre d'utiliser le logiciel R pour **décrire** et **représenter** des **données relatives à l'enquête nationale sur la demande en Habitat**.

2. Structure de données relatives à l'enquête nationale sur la demande en Habitat

Chaque ménage est décrit par un ensemble de caractéristiques appelées variables. (Voir l'exemple ci-après) :

	A	B	C	D	E	F	G	H	I
1	MILIEU_(	REGION_(	PROV_PREF_(	COM_VILLE_(	CODE_DISTRICT_(	NUM_LOGEM_DISTRICT_(	NBR_MENAGE_HABIT_(	NUM_MENAGE_ECH_(	NUM_INTERESSES_ACHAT_IMMOC_01_Q5
2	URBAIN	FES-MEKNES	451	109	145	97	1	33	2
3	URBAIN	MARRAKESH	211	105	10	26	1	7	1
4	URBAIN	DRAA-TAFIL	201	105	113	37	1	13	1
5	URBAIN	TANGER-TET	511	101	20	7	1	3	2
6	URBAIN	MARRAKESH	351	101	3	9	1	3	1
7	URBAIN	MARRAKESH	211	105	10	14	1	4	1
8	URBAIN	SOUSS-MASS	273	105	91	28	1	14	1
9	URBAIN	MARRAKESH	211	105	10	90	1	23	1
10	URBAIN	EDDAKHLA-(	391	101	61	79	1	29	1
11	URBAIN	GRAND CASA	117	103	1	11	1	6	1
12	URBAIN	DRAA-TAFIL	201	105	113	43	1	15	1
13	URBAIN	RABAT-SALE	441	103	271	80	1	27	1
14	URBAIN	LAAYOUNE-S	321	103	85	44	1	22	1
15	URBAIN	EDDAKHLA-(	391	101	42	33	6	7	1
16	URBAIN	RABAT-SALE	441	103	271	57	1	20	1
17	URBAIN	EDDAKHLA-(	391	101	42	93	1	19	1
18	URBAIN	ORIENTAL	381	105	68	11	1	11	1
19	URBAIN	GRAND CASA	117	103	1	45	1	23	1
20	URBAIN	ORIENTAL	381	107	29	41	1	14	1
21	URBAIN	ORIENTAL	381	107	29	53	1	18	1

Exemple de feuille de données relatives à l'enquête nationale sur la demande en Habitat

Remarque : ces données factices sont générées par le la plateforme CAPI

Ce jeu artificiel de données relatives à l'enquête nationale sur la demande en Habitat est enregistré dans un fichier Excel **« data_habitat.xls »**. Ce dernier est composé de sept feuilles de données :

1. Localisation_geograph_menage
2. Caracteristiques_membres_menage
3. Caracteristiques_chef_menage
4. Mobilte_menage_10annees
5. Condition_habitat
6. Degre_satisfaction

7. Demandeur_habitat

3. Installation des logiciels R et RStudio

1. Rendez-vous ici : https://cran.rstudio.com/bin/windows/base/R-3.2.2-win.exe
2. Cliquez sur : « R-3.2.2-win.exe ».
3. D'abord, enregistrez le fichier nommé «R-3.2.2-win.exe » sur votre disque.
4. Ensuite, double cliquez sur ce fichier et suivez les consignes d'installation en laissant l es valeurs par défaut.
5. Enfin, si tout a fonctionné vous devriez avoir la possibilité de lancer le programme de puis le menu « démarrer » de Windows.

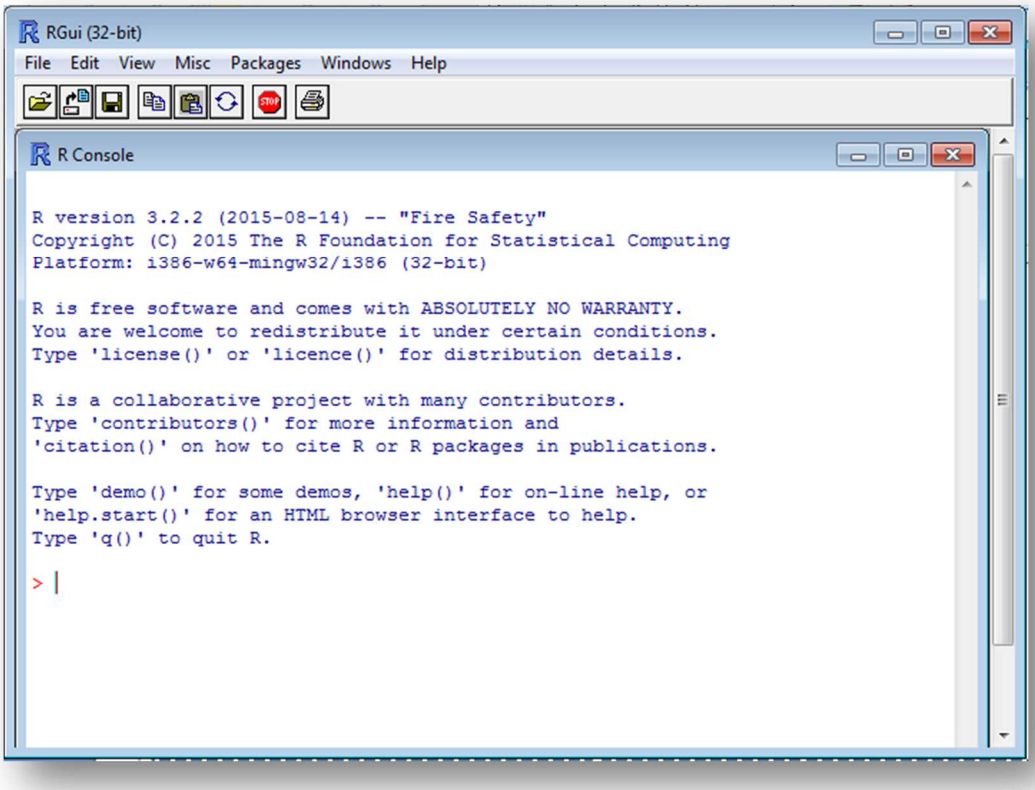


Figure :

6. En suivant les mêmes étapes, téléchargez et installez le fichier nommé RStudio-0.99. 484 sur https://download1.rstudio.org/RStudio-0.99.486.exe

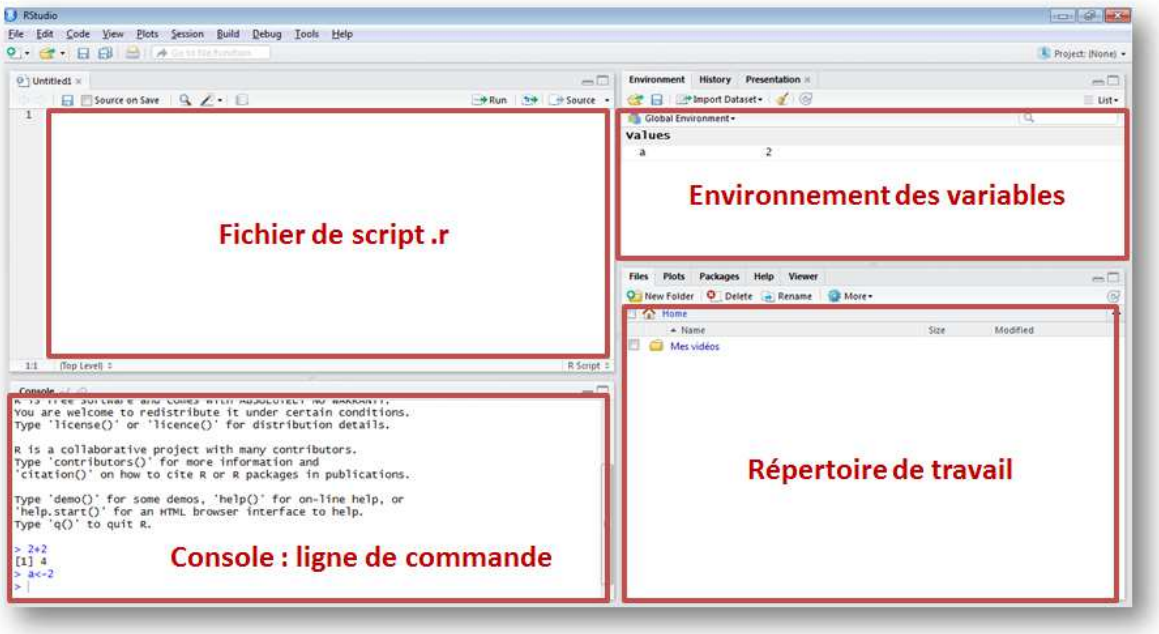


Figure :

4. Démarrer le logiciel RStudio

Avec Windows : menu démarrer → programmes → RStudio

5. Chargement des données dans le logiciel RStudio

Pour pouvoir charger les données de l'enquête de puis un fichier de type *.xls vers le logiciel R, vous devriez suivre les étapes suivantes :

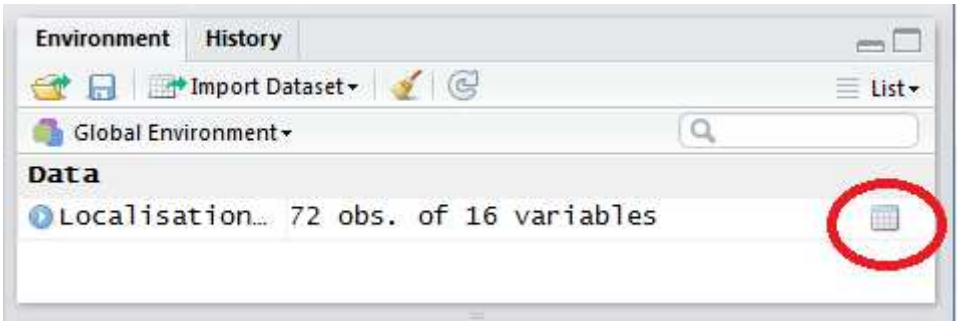
1. Copier le fichier *.xls dans votre répertoire de travail du logiciel RStudio.
   Remarque : vous pouvez savoir le chemin du répertoire de travail à partir de la comm ande suivante : `> getwd()`
   `[1] "C:/Users/imade/Documents"`
2. Installation de packages : `> install.packages("gdata")`
3. Chargement de la bibliothèque : `> library(gdata)`
4. Chargement de données : `> Localisation_geograph_menage<-read.xls(xls = "data_habitat.xls",sheet=1)`

Remarque : sheet=1 signifie la première feuille du classeur `data_habitat.xls`

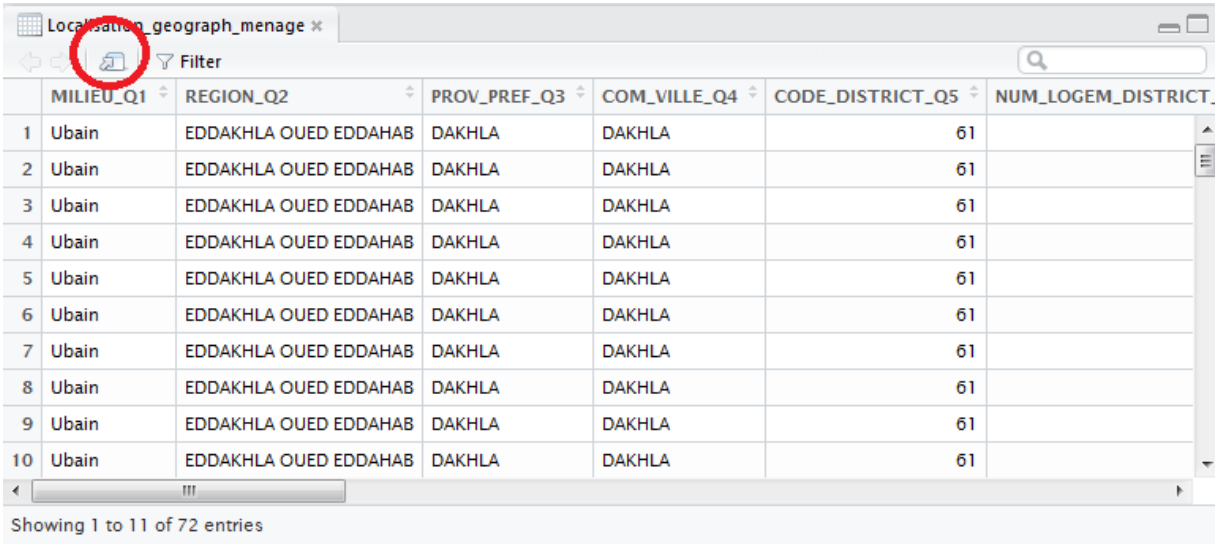
Pour les fichiers de type xlsx, vous devrez suivre les étapes suivantes :

1. Copier le fichier *.xlsx dans votre répertoire de travail dans le logiciel RStudio.
   Remarque : vous pouvez savoir le chemin du répertoire à partir de la commande suiv ante : `> getwd()`
   `[1] "C:/Users/imade/Documents"`
2. Installation de packages : `> install.packages("xlsx ")`
3. Chargement de la bibliothèque : `> library(xlsx)`
4. Chargement de données : `> Localisation_geograph_menage<-read.xlsx(xlsx = "data_habitat.xlsx",sheet=1)`

Une fois les données sont chargées, vous pourriez les visualiser en cliquant sur le bouton ci-dessus :



La fenêtre suivante sera affichée :



	MILIEU_Q1	REGION_Q2	PROV_PREF_Q3	COM_VILLE_Q4	CODE_DISTRICT_Q5	NUM_LOGEM_DISTRICT_
1	Ubain	EDDAKHLA OUED EDDAHAB	DAKHLA	DAKHLA	61	
2	Ubain	EDDAKHLA OUED EDDAHAB	DAKHLA	DAKHLA	61	
3	Ubain	EDDAKHLA OUED EDDAHAB	DAKHLA	DAKHLA	61	
4	Ubain	EDDAKHLA OUED EDDAHAB	DAKHLA	DAKHLA	61	
5	Ubain	EDDAKHLA OUED EDDAHAB	DAKHLA	DAKHLA	61	
6	Ubain	EDDAKHLA OUED EDDAHAB	DAKHLA	DAKHLA	61	
7	Ubain	EDDAKHLA OUED EDDAHAB	DAKHLA	DAKHLA	61	
8	Ubain	EDDAKHLA OUED EDDAHAB	DAKHLA	DAKHLA	61	
9	Ubain	EDDAKHLA OUED EDDAHAB	DAKHLA	DAKHLA	61	
10	Ubain	EDDAKHLA OUED EDDAHAB	DAKHLA	DAKHLA	61	

Showing 1 to 11 of 72 entries

Le bouton encerclé vous permettra de zoomer sur les données

Atelier : En suivant les mêmes étapes, charger toutes les données « artificielles » relatives à l'enquête.

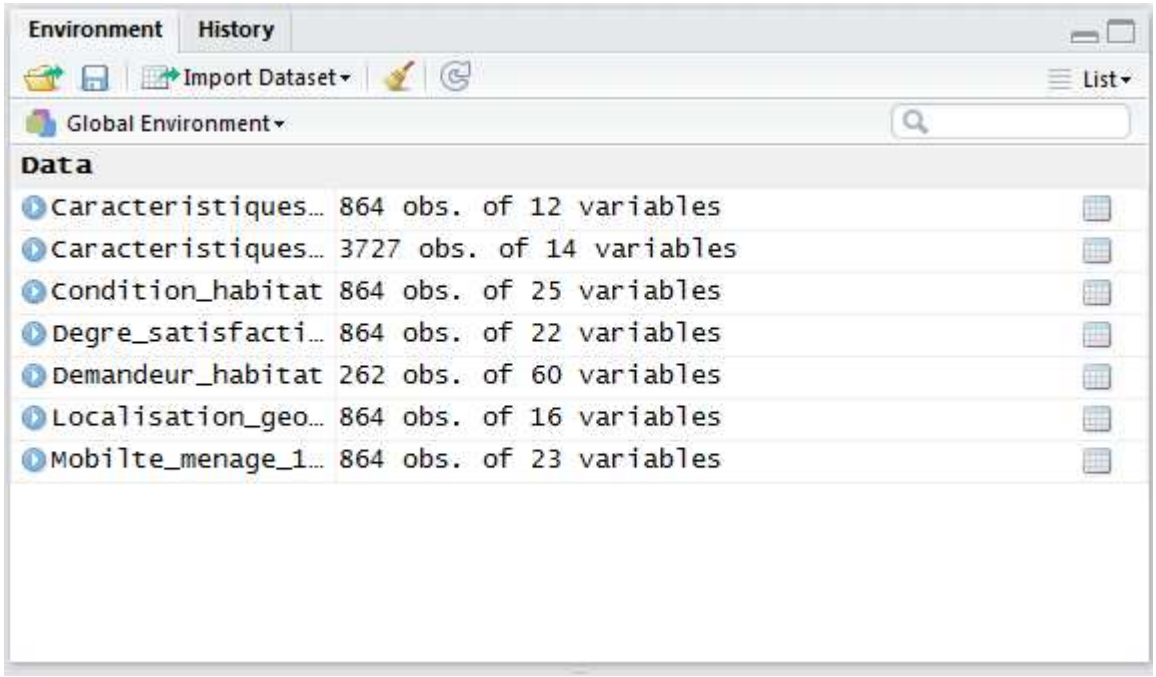
Solution :

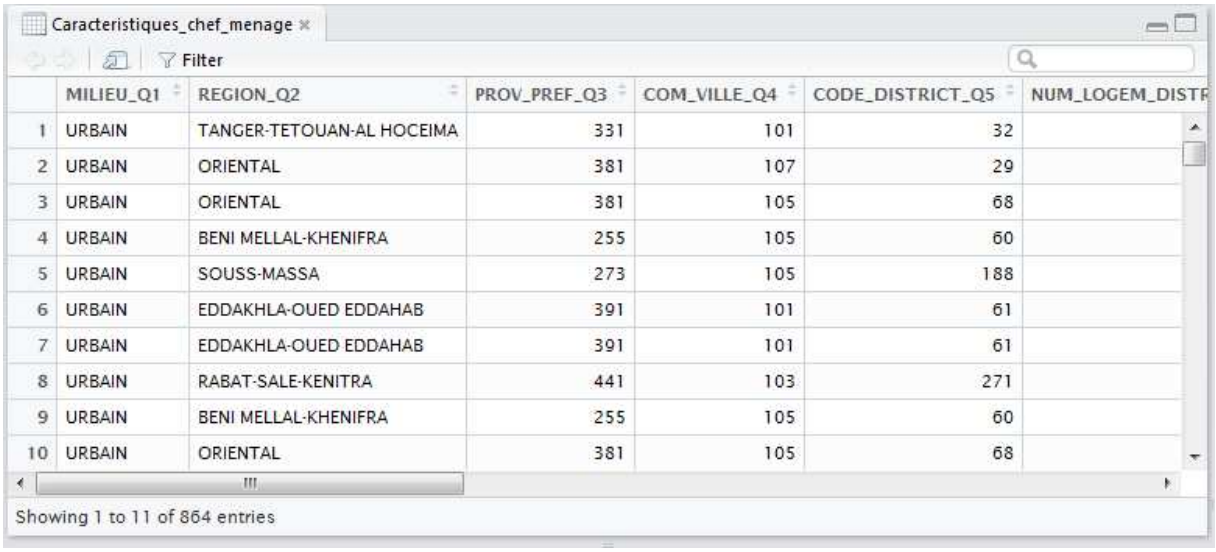
```
> Localisation_geograph_menage<-read.xls(xls = "data_habitat.xls",sheet=1)
> Caracteristiques_membres_menage<-read.xls(xls = "data_habitat.xls",sheet=
2)
> Caracteristiques_chef_menage<-read.xls(xls = "data_habitat.xls",sheet=3)
> Mobilte_menage_10annees<-read.xls(xls = "data_habitat.xls",sheet=4)
> Condition_habitat<-read.xls(xls = "data_habitat.xls",sheet=5)
> Degre_satisfaction<-read.xls(xls = "data_habitat.xls",sheet=6)
> Demandeur_habitat<-read.xls(xls = "data_habitat.xls",sheet=7)
```

Remarque : Le pire ennemi du statisticien, tous les enquêteurs le savent, est la valeur manquante. En R, les valeurs manquantes sont codées NA ou <NA>. Il faut donc préciser à R le type de valeur manquante qu'il va rencontrer dans le fichier. Cela se fait en ajoutant l'option na.string="codage_Manquante" dans la ligne de lecture.

```
read.xls("nom_de_fichier.xls", na.string = "")
```

Vous devez avoir le résultat suivant :





Caracteristiques_chef_menage

Filter

	MILIEU_Q1	REGION_Q2	PROV_PREF_Q3	COM_VILLE_Q4	CODE_DISTRICT_Q5	NUM_LOGEM_DISTR
1	URBAIN	TANGER-TETOUAN-AL HOCEIMA	331	101	32	
2	URBAIN	ORIENTAL	381	107	29	
3	URBAIN	ORIENTAL	381	105	68	
4	URBAIN	BENI MELLAL-KHENIFRA	255	105	60	
5	URBAIN	SOUSS-MASSA	273	105	188	
6	URBAIN	EDDAKHLA-OUED EDDAHAB	391	101	61	
7	URBAIN	EDDAKHLA-OUED EDDAHAB	391	101	61	
8	URBAIN	RABAT-SALE-KENITRA	441	103	271	
9	URBAIN	BENI MELLAL-KHENIFRA	255	105	60	
10	URBAIN	ORIENTAL	381	105	68	

Showing 1 to 11 of 864 entries

Remarque : les données sont actuellement sous forme d'un tableau de données « data.frame »

Pour afficher les noms des colonnes du tableau de données « Degre_satisfaction » :

```
> colnames(Degre_satisfaction)
 [1] "MILIEU_Q1"                          "REGION_Q2"
 [3] "PROV_PREF_Q3"                       "COM_VILLE_Q4"
 [5] "CODE_DISTRICT_Q5"                   "NUM_LOGEM_DISTRICT_Q6"
 [7] "NBR_MENAGE_HABIT_Q7"                "NUM_MENAGE_ECH_Q8"
 [9] "SATISFAIT_CONDITIONS_HABITATIOP_Q44" "RAISONS_N_SATISFACTION1_Q45"
[11] "AUTRE_N_SATIS1_Q45"                 "RAISONS_N_SATISFACTIOO2_Q45"
[13] "AUTRE_N_SATIU2_Q45"                 "RAISONS_N_SATISFACTIOP3_Q45"
[15] "AUTRE_N_SATIT3_Q45"                 "PERE_RECHERCHE_HABIT_Q46"
[17] "PERE_5_AN_AVENIR_Q47"               "PERE_AN_RECHERCHE_Q48"
[19] "RAISON_NON_ACHAT_IMMOC_Q49"         "AUTRE_RAISON_N_ACHAT_Q49"
[21] "FAM_RECHERCHE_Q50"                  "NBR_DESIR_ACHAT_Q51"
```

Pour afficher la dimension du tableau de données « Degre_satisfaction »:

```
> dim(Degre_satisfaction)
[1] 864  22
```

Pour afficher les 6 premières lignes :

```
> head(Degre_satisfaction)
  MILIEU_Q1                 REGION_Q2 PROV_PREF_Q3 COM_VILLE_Q4 CODE_DISTRI
CT_Q5
1    URBAIN TANGER-TETOUAN-AL HOCEIMA          511          101
20
2    URBAIN TANGER-TETOUAN-AL HOCEIMA          331          101
32
3    URBAIN TANGER-TETOUAN-AL HOCEIMA          331          101
32
```

```
4     URBAIN TANGER-TETOUAN-AL HOCEIMA          331           101
32
5     URBAIN TANGER-TETOUAN-AL HOCEIMA          331           101
32
6     URBAIN TANGER-TETOUAN-AL HOCEIMA          331           101
32
```

Pour afficher les 6 dernières lignes :

```
> tail(Degre_satisfaction)
    MILIEU_Q1               REGION_Q2 PROV_PREF_Q3 COM_VILLE_Q4 CODE_DISTRI
CT_Q5
859    URBAIN GRAND CASABLANCA-SETTAT          141           101
290
860    URBAIN   EDDAKHLA-OUED EDDAHAB          391           101
42
861    URBAIN       GUELMIM-OUED NOUN          261           103
188
862    URBAIN       GUELMIM-OUED NOUN          261           103
39
863    URBAIN   EDDAKHLA-OUED EDDAHAB          391           101
61
864    URBAIN   EDDAKHLA-OUED EDDAHAB          391           101
61
```

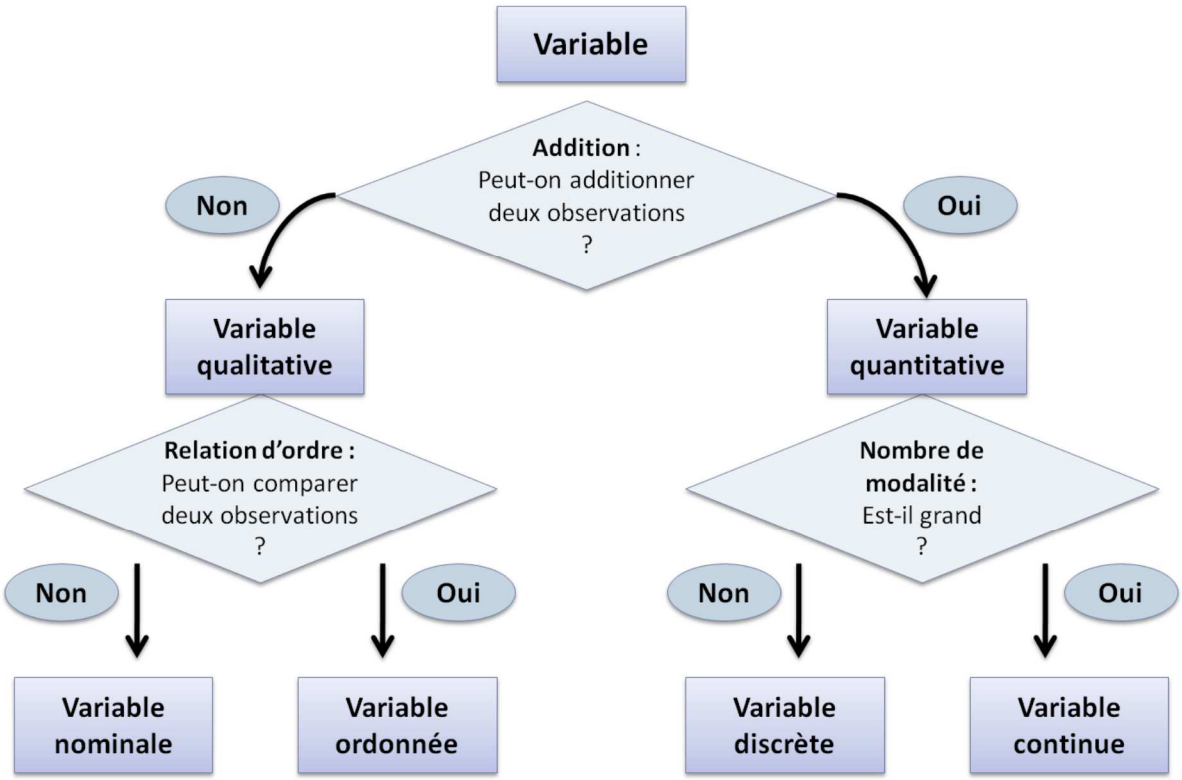
6. Manipulation d'un data.frame

```
> ### Deuxième colonne
> Degre_satisfaction [ ,2]
> ### Colonne milieu
> Degre_satisfaction$MILIEU_Q1
> ### Troisième ligne
> Degre_satisfaction [3 ,]
> ### Affichage d' une valeur précise
> Degre_satisfaction [3 ,2]
> Degre_satisfaction$ MILIEU_Q1 [3]
```

Médication d'une valeur :

```
> ### Modification de la cinquième valeur
> Degre_satisfaction$MILIEU_Q1[5] <- 34
```

7. Variable et modalité



Une variable est dite qualitative si elle n'est pas quantifiable : **REGION_Q2**

Une variable est dite quantitative si elle est quantifiable : **NBR_CHAMBRE_Q34**

On appelle modalités les différentes valeurs qui caractérisent une variable. Par exemple, dans un questionnaire, une réponse à une question est une modalité.

Une variable quantitative est dite discrète si ses modalités prennent des valeurs positives et entières : **NBR_CHAMBRE_Q34**

Une variable quantitative est dite continue si ces modalités sont mesurables : **SUPERFICIE_UTILISE_MA_Q35**

Une variable qualitative est dite ordinale si on peut classer ses modalités : **SATISFAIT_CONDITIONS_HABITATIOP_Q44 (Satisfait, Non satisfait, Pas satisfait du tout …)**

Sinon elle est dite nominale : **SEXF_01_Q18**

8. Type de variable sous R

Chaque colonne correspond à une variable et a donc un type. Les différents types de variables statistique correspondent aux types R suivant :

En statistique	Sous R
Nominale	`factor`
Ordonnée	`ordered`
Discrète	`numeric (ou integer)`
Continue	`numeric (ou integer)`

```
> ### Le type des colonnes
> str(Degre_satisfaction)
```

Transtypage

Pour transformer une variable numérique en facteur, il faut utiliser la fonction `as.factor`. `as.factor(Degre_satisfaction$CODE_DISTRICT_Q5)` permet de considérer la colonne non plus comme une variable numérique mais comme une nominale.

9. Population, échantillon, individu

On appelle **population** un ensemble d'éléments de même nature : **Les ménages**.

Comme il est souvent difficile d'étudier une population dans son intégralité on travaille alors sur une partie de la population que l'on appelle échantillon. Un **échantillon** est donc un sous-ensemble d'une population. Constituer un échantillon c'est effectuer un sondage.

On appelle **individu** un élément appartenant à une population ou à un échantillon : **Un ménage**

10. La statistique

La statistique constitue un ensemble de procédures destinées à la collecte, l'organisation, la synthèse, l'analyse et la représentation de données.

- On parle de statistique descriptive lorsque l'on décrit et analyse des données sans en généraliser les conclusions.
- On parle de statistique inférentielle lorsque l'on cherche à généraliser à d'autres ensembles les caractéristiques d'un échantillon.

11. Statistique descriptive

L'objet de la statistique descriptive est de synthétiser des informations contenues dans un jeu de données.

Résumer les informations issues des variables quantitatives est une étape importante dans un travail d'analyse de donnée. La synthèse de ces données se fait notamment grâce à des résumés numériques. Nous allons voir dans ce module comment résumer une série d'observations par des indicateurs caractérisant la distribution.

Deux types d'indicateurs seront abordés :

- deux caractéristiques de tendance centrale qui sont représentatives d'une distribution statistique (moyenne et médiane),

- deux caractéristiques de dispersion qui témoignent de la dispersion d'une distribution autour d'une valeur centrale (variance et écart type).

12. L'analyse univariée

Permet de mieux appréhender une variable. Elle comporte quatre étapes :

1. Calcul des effectifs
2. Calcul de la centralité
3. Calcul de la dispersion
4. Représentation graphique

Ces étapes varient selon le type de variable. Voilà le détail des étapes en fonction du type de variable :

Étape	Nominale	Ordonnée	Discrète	Continue
1. Effectifs	A faire	A faire	A faire	Inutile
2. Centralité	Mode	Médiane	Moyenne et Médiane	Moyenne et Médiane
3. Dispersion	N'existe pas	Quartile	Écart type et quartiles	Écart type et quartile
4. Graphique	Histogramme des effectifs	Histogramme des effectifs	Histogramme des effectifs, boîte à moustache	Distribution et boîte à moustache

13. Effectifs

Les effectifs se calculent pour les variables nominales, ordonnée et discrètes. Cela se fait grâce à l'instruction `table()` :

Remarque : effectifs se calcul pour les variables de type factor et ordered.

Effectifs de genre :

```
> table(Caracteristiques_membres_menage$SEXF_01_Q18)

   1    2
1924 1801
```

Remarque : 1 pour Homme et 2 pour Femme.

Effectifs des niveaux :

```
> table(Caracteristiques_chef_menage$NIV_ETUE_Q21)

Collégial Formation professionnelle        Lycéen
112       10                                  121
NSP       Primaire             Quoranique
3         181                  64
Sans                 Supérieur
270                  103
```

14. Résumés de variables quantitatives : les caractéristiques de position

Calculer une caractéristique de tendance centrale revient à définir une valeur autour de laquelle se répartissent des observations.

Le mode

Le mode s'obtient par lecture de la table des effectifs en prenant le plus grand. Si les modalités sont très nombreuse, on peut trier les effectifs avec l'instruction sort de manière décroissante en utilisant l'option decreasing=TRUE (afin que le mode soit en tête).

```
MyMOde <- function(x){
  sort(table(x),decreasing = TRUE)[1]
}
```

La moyenne arithmétique

$$\bar{x} = \frac{\sum x_i}{n}$$

Pour calculer la moyenne arithmétique sur R :

```
> mean(Demandeur_habitat$PRIX_MAX_ACHAT_01_Q70)
[1] NA
> mean(Demandeur_habitat$PRIX_MAX_ACHAT_01_Q70,na.rm = T)
[1] 416529.4
> mean(Demandeur_habitat$PRIX_MAX_CREDIT_01_Q73,na.rm = T)
[1] 1576.471
```

Remarque : la moyenne arithmétique est très sensible aux variations des valeurs extrêmes, c'est pourquoi c'est une caractéristique de tendance centrale considérée comme peu robuste. Pour compléter les informations qu'elle fournit sur la distribution on fera donc aussi appelle à la médiane.

La médiane

La médiane est la valeur de la variable qui divise l'échantillon en deux ensembles de mêmes effectifs.

Pour calculer la médiane sous R :

```
> median(Demandeur_habitat$PRIX_MAX_ACHAT_01_Q70,na.rm = T)
[1] 5e+05
> median(Demandeur_habitat$PRIX_MAX_CREDIT_01_Q73,na.rm = T)
[1] 1300
```

Médiane d'une ordonnée :

La médiane d'une variable ordonnée n'est pas calculée automatiquement par R. Il faut donc le faire manuellement. Pour cela, trois étapes :

1. Calcul du rang de la médiane (après exclusion des manquantes).
2. Ordonnancement de la variable
3. Combinaison de 1 et 2, sélection la modalité du milieu

La médiane des niveaux d'études des chefs de ménages :

```
> MyMedian<-function(x){
+ sort(x)[round((length(na.omit(x))+1)/2)]
+ }
> MyMedian(Caracteristiques_chef_menage$NIV_ETUE_Q21)
[1] Quoranique
8 Levels: Collégial Formation professionnelle Lycéen NSP Primaire Quoraniqu
e ... Supérieur
```

15. Résumés de variables quantitatives : les caractéristiques de dispersion

La dispersion est une notion clé en statistique. Les caractéristiques de dispersion témoignent de la répartition de la distribution autour d'une valeur centrale. Ce sont des éléments très importants qui décrivent comment les valeurs de la variable s'étalent autour de la moyenne ou de la médiane. La variance et l'écart type sont les deux mesures les plus fréquemment utilisées.

Remarque : la moyenne, l'écart type et le quartile se calculent pour le type « numeric »

Les quartiles, les déciles et les centiles

Il s'agit de valeurs d'une série ou d'une distribution statistique rangée dans un ordre particulier (croissant ou décroissant), partageant l'effectif total en plusieurs parties égales. Les quartiles sont dénombrables au nombre de 3 ($Q1$,$Q2$,$Q3$: $Q1$ représentant le premier quart, $Q2$, le second, et $Q3$, le troisième), les déciles, au nombre de 9 (puisque qu'il divise la série en 10 parties égales) et les centiles, au nombre de 99 (puisqu'il divise la série en 100 parties égales)

Pour calculer les quartiles sous R :

```
> quantile(Demandeur_habitat$PRIX_MAX_CREDIT_01_Q73,na.rm = T)
   0%   25%   50%   75%  100%
  500  1000  1300  1700 15000
```

Pour une variable ordonnée, la méthode est la même que pour la médiane.

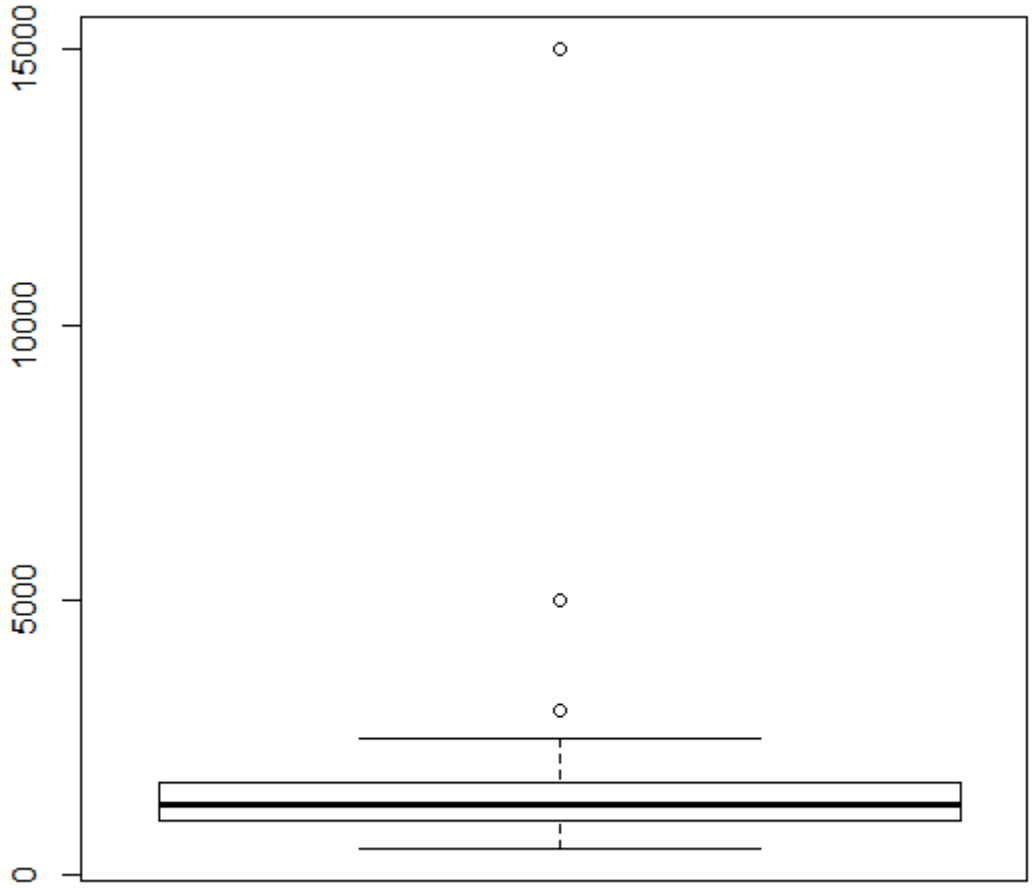
```
> MyQuartiles<- function(x){
+ levels(x)[quantile(as.numeric(x),na.rm=TRUE)]
+ }
> MyQuartiles(Caracteristiques_chef_menage$NIV_ETUE_Q21)
[1] "Collégial"  "Lycéen"     "Quoranique" "Sans"       "Supérieur"
```

Résumé des caractéristiques

```
> summary(Demandeur_habitat$PRIX_MAX_CREDIT_01_Q73)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    500    1000    1300    1576    1700   15000     177
```

Ces valeurs peuvent être résumées graphiquement :

```
> boxplot(Demandeur_habitat$PRIX_MAX_CREDIT_01_Q73)
```



Remarques : nous allons s'intéresser par la suite à la représentation graphique de données.

Variance et écart type

L'écart type mesure la dispersion d'une distribution autour de la moyenne. En quelque sorte, l'écart type évalue la « largeur moyenne » de la distribution, il s'exprime donc dans la même unité que la variable.

Plus l'écart-type est faible, plus la distribution est homogène. Autrement dit, plus l'écart type est petit, plus les valeurs de la distribution sont proches les unes des autres. À l'inverse, plus l'écart type est grand, plus la distribution est étalée : plus les valeurs sont éloignées les unes des autres.

La variance

Pour calculer l'écart-type d'une distribution il faut auparavant calculer la variance. En effet, l'écart-type est la racine carré de la variance. La variance est noté s^2 tant dis que l'écart-type est noté s .

Pour calculer la variance il existe un moyen mnémotechnique : « la variance est la moyenne des carrés moins le carré de la moyenne ».

Voyons ce que cela veut dire.

La moyenne des carrés

Que signifie "la moyenne des carrés" ?

Les carrés sont les valeurs de la variable x_i élevées au carré (multiplié par eux-mêmes), donc $x_i \times x_i$ noté x_i^2 .

Pour trouver la moyenne des carrés il suffit d'appliquer la formule de la moyenne : la somme de la distribution des carrés, divisée pas le nombre de termes soit $\sum x_i^2$ divisé par n donc :

$$\frac{\sum x_i^2}{n}$$

Voilà pour "la moyenne des carrés".

Le carré de la moyenne

Voyons maintenant ce que signifie "le carré de la moyenne"

C'est simplement la moyenne de la distribution multipliée par elle-même :

$(\frac{\sum x_i}{n}) \times (\frac{\sum x_i}{n})$ soit $(\frac{\sum x_i}{n})^2$

La moyenne des carrés moins le carré de la moyenne

Reprenons donc notre moyen mnémotechnique :

« la variance (s^2) est la moyenne des carrés ($\frac{\sum x_i^2}{n}$) moins le carré de la moyenne $(\frac{\sum x_i}{n})^2$ » la formule est donc :

$$s^2 = \frac{\sum x_i^2}{n} - (\frac{\sum x_i}{n})^2$$

L'écart type

L'écart-type est la racine carrée de la variance que l'on va donc noter ainsi :

$$s = \sqrt{\frac{\sum x_i^2}{n} - (\frac{\sum x_i}{n})^2}$$

L'écart type et la variance se calculent respectivement à l'aide de sd() et var(), avec l'option na.rm=TRUE pour supprimer les manquantes :

```
> var(Demandeur_habitat$NBR_CHMBRES_01_Q64)
[1] NA
> var(Demandeur_habitat$NBR_CHMBRES_01_Q64,na.rm = T)
[1] 1.072389

> sd(Demandeur_habitat$NBR_CHMBRES_01_Q64,na.rm = T)
[1] 1.035562
```

MAD : Median Absolute Deviation (un estimateur robuste)

Données aberrantes ou outliers : sont des observations atypiques bien éloignées de la masse des données et sont des points isolés ou en petit groupes de points ; dues à des erreurs de copie, de calcul, de changement d'unités, ou des données n'obéissant pas au même modèle (présence de plusieurs classes). Les données aberrantes sont les plus dangereuses pour l'estimateur.

MAD : définit la variation de l'écart absolu à la médiane. Il est moins affecté par les données extrêmes. On l'utilise comme estimateur de l'écart-type.

Exprime l'exactitude dans les mêmes unités que les données, ce qui aide à conceptualiser l'importance de l'erreur.

Sous R la MAD est calculé par la fonction `mad()` :

```
> mean(Demandeur_habitat$PRIX_MAX_CREDIT_01_Q73,na.rm = T)
[1] 1576.471
> median(Demandeur_habitat$PRIX_MAX_CREDIT_01_Q73,na.rm = T)
[1] 1300
> mad(Demandeur_habitat$PRIX_MAX_CREDIT_01_Q73,na.rm = T)
[1] 444.78
> mad(Demandeur_habitat$PRIX_MAX_CREDIT_01_Q73,na.rm = T,center = median(De
mandeur_habitat$PRIX_MAX_CREDIT_01_Q73,na.rm = T))
[1] 444.78
> mad(Demandeur_habitat$PRIX_MAX_CREDIT_01_Q73,na.rm = T,center = mean(Dema
ndeur_habitat$PRIX_MAX_CREDIT_01_Q73,na.rm = T))
[1] 854.6753
> mad(Demandeur_habitat$PRIX_MAX_CREDIT_01_Q73,na.rm = T, constant = 1.483)
[1] 444.9
> mad(Demandeur_habitat$PRIX_MAX_CREDIT_01_Q73,na.rm = T, constant = 2)
[1] 600
```

Remarque : Souvent on utilise comme valeur de s : s = 1.483MAD (1.483 vient d'une hypothèse que les données sont gaussiennes)

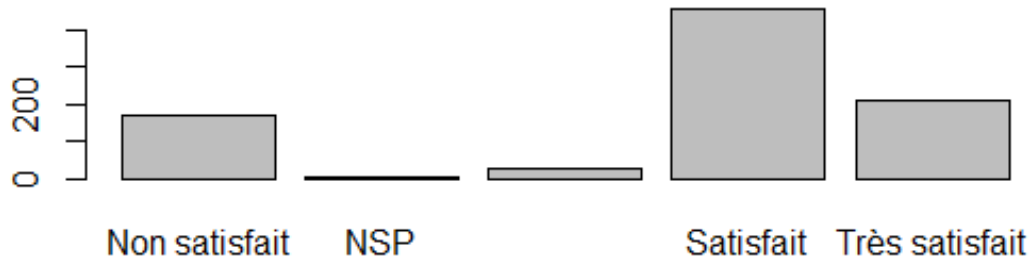
16. Représentation graphique

R dispose d'un grand nombre d'outils graphiques permettant de représenter des données. Là encore, la représentation graphique dépend du type de variable.

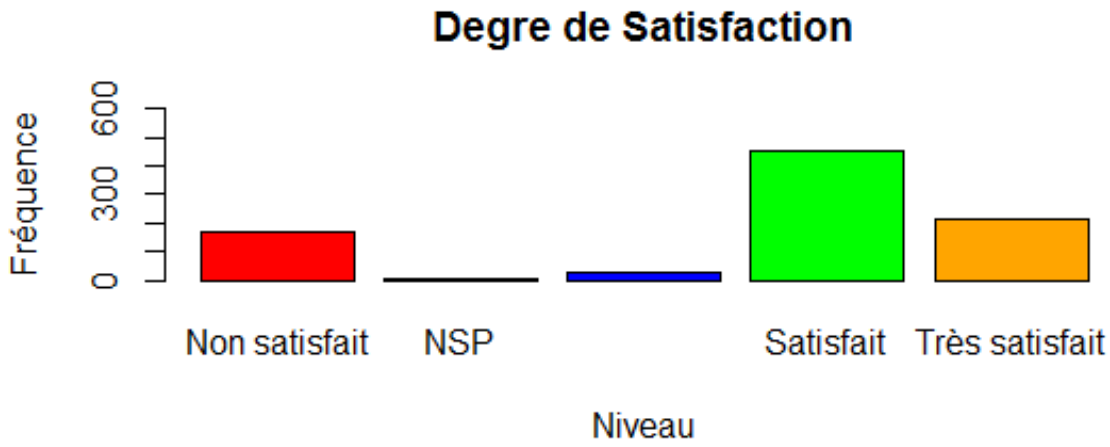
Diagramme en baton

Pour les variables pour lesquelles il est possible de calculer les effectifs, on peut tracer un diagramme en baton :

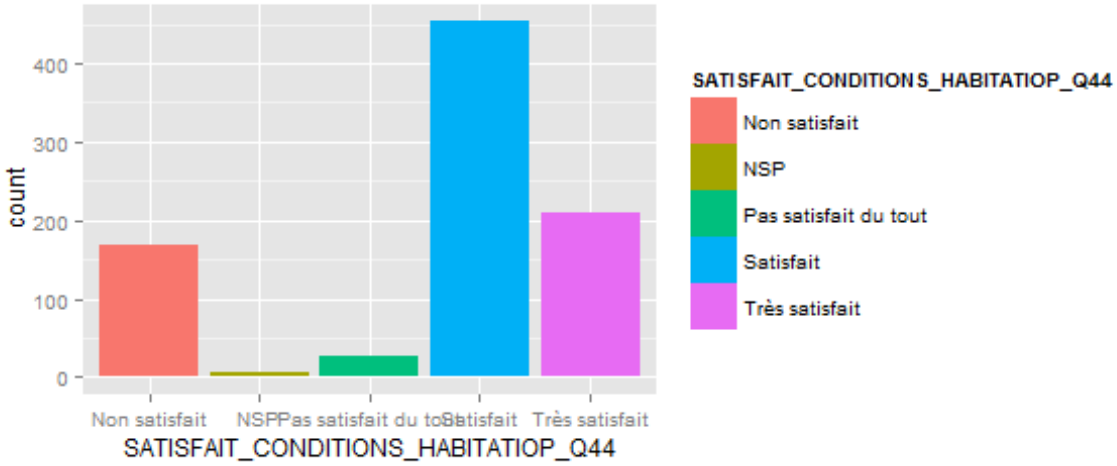
```
> barplot(table(Degre_satisfaction$SATISFAIT_CONDITIONS_HABITATIOP_Q44))
```



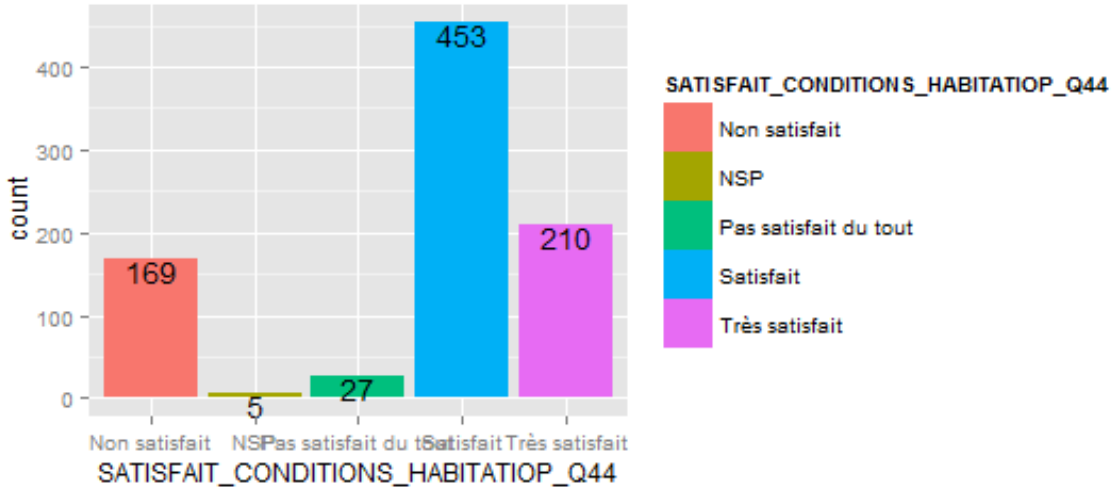

```
> colors<-c("red","yellow","blue","green","orange","cyan")
> barplot(table(Degre_satisfaction$SATISFAIT_CONDITIONS_HABITATIOP_Q44), ylim = c(0,600),main = "Degre de Satisfaction",xlab = "Niveau",ylab = "Fréquence",col=colors, legend.text = F,horiz = F)
```



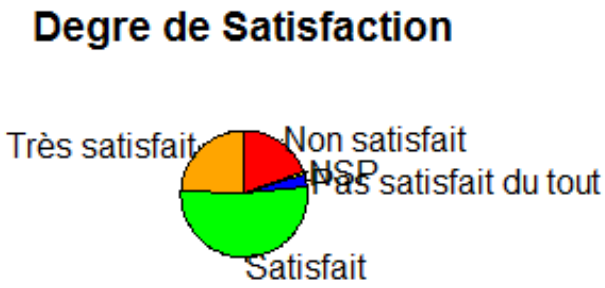

```
> library(ggplot2)
> qplot(SATISFAIT_CONDITIONS_HABITATIOP_Q44, data=Degre_satisfaction, geom="bar",fill=SATISFAIT_CONDITIONS_HABITATIOP_Q44)
```



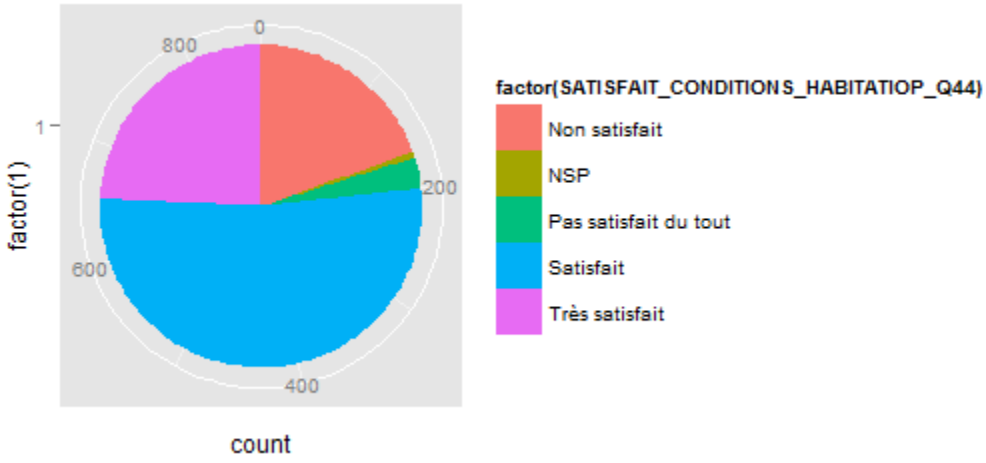

```
> ggplot(Degre_satisfaction, aes(x=SATISFAIT_CONDITIONS_HABITATIOP_Q44,fill=SATISFAIT_CONDITIONS_HABITATIOP_Q44)) + geom_bar()+geom_text(stat='bin',aes(label=..count..),vjust=1)
```



Il est également possible de tracer des camemberts, mais cette représentation graphique est fortement déconseillée, l'œil humain n'étant en effet pas adapté à l'évaluation des distances angulaires. Néanmoins, cela peut se faire avec pie (fortement déconseillé).



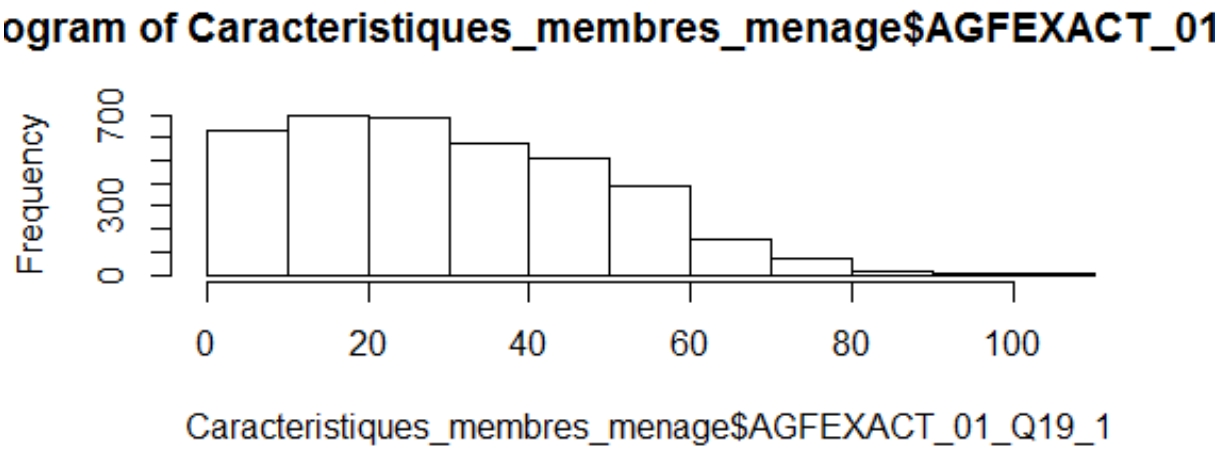

```
> ggplot(Degre_satisfaction, aes(x = factor(1), fill = factor(SATISFAIT_CON
DITIONS_HABITATIOP_Q44))) +
+     geom_bar(width = 1)+ coord_polar(theta = "y")
```



Histogramme

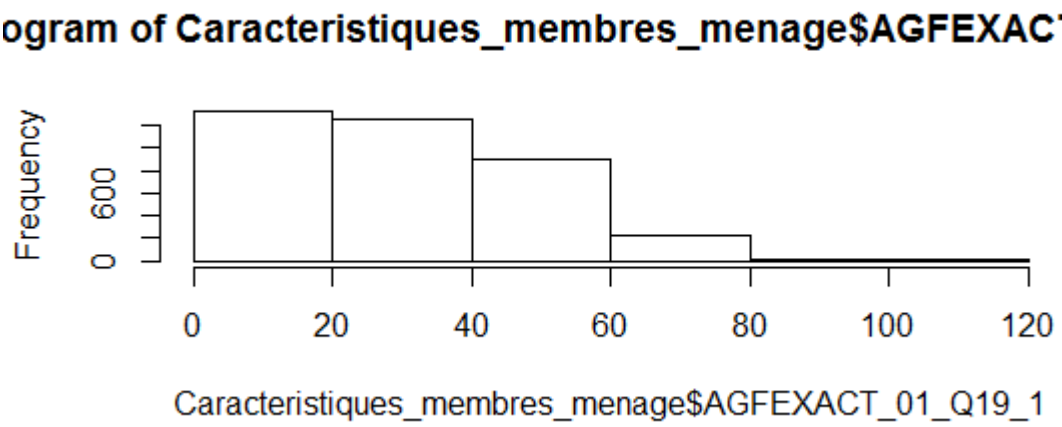
Un diagramme en baton est une représentation graphique adaptée aux variables ayant peu de modalités. Si on l'utilise sur une variable continue, on obtient un graphique peut informatif et supprimant les modalités manquantes :

```
> hist(Caracteristiques_membres_menage$AGFEXACT_01_Q19_1)
```

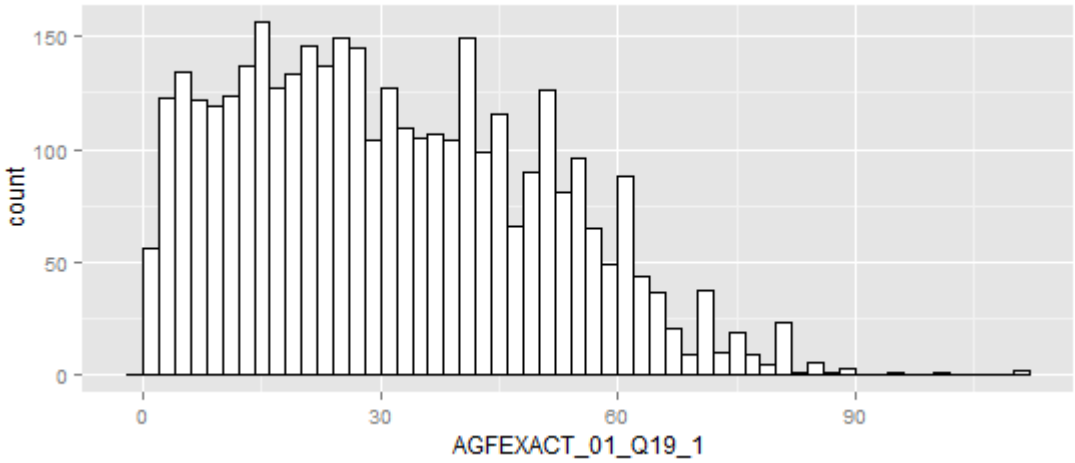



```
>barplot(Caracteristiques_membres_menage$AGFEXACT_01_Q19_1)
```

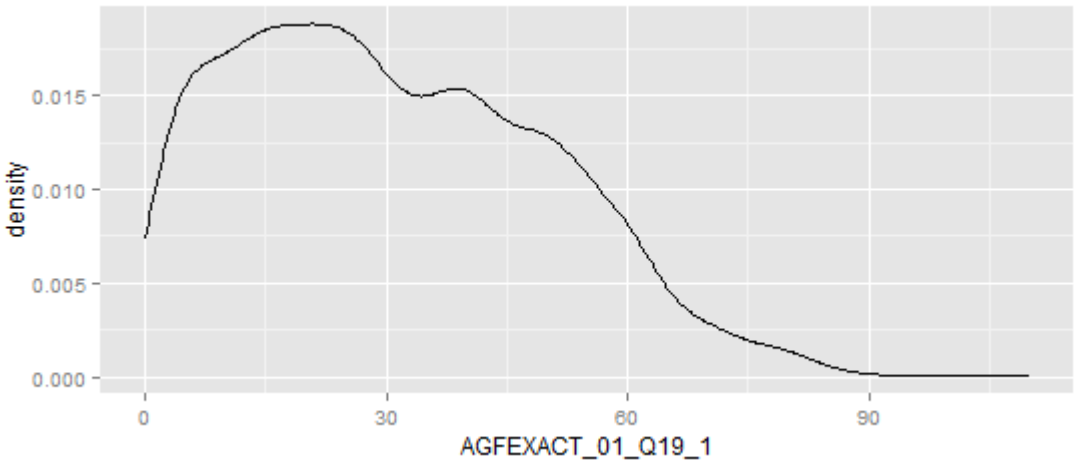

```
> hist(Caracteristiques_membres_menage$AGFEXACT_01_Q19_1,breaks=5)
```




```
> ggplot(Caracteristiques_membres_menage, aes(x=AGFEXACT_01_Q19_1)) +
+     geom_histogram(binwidth=2, fill="white", colour="black")
```



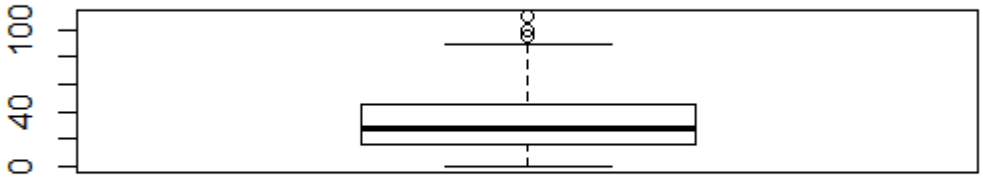

```
> ggplot(Caracteristiques_membres_menage, aes(x=AGFEXACT_01_Q19_1)) + geom_
line(stat="density")
```



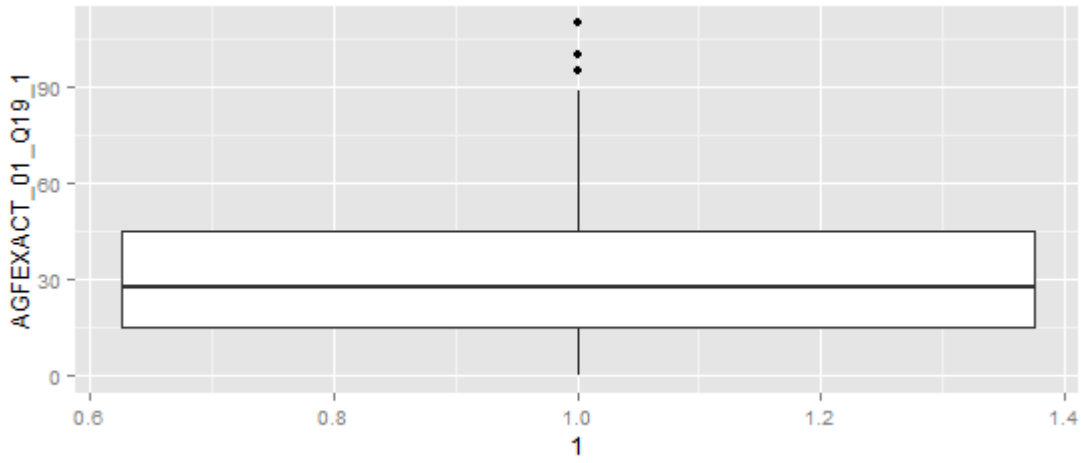
Boîte à moustaches

Les boîtes à moustaches sont des représentations graphiques utilisables pour des variables numériques. La boîte est délimitée par le premier et les troisièmes quartiles, elle contient donc 50% de la population. Les moustaches encadrent les individus proches du centre. Au-delà des moustaches, on trouve soit les valeurs aberrantes, soit les valeurs éloignées du centre (valeurs extrêmes).

```
> boxplot (Caracteristiques_membres_menage$AGFEXACT_01_Q19_1)
```




```
> ggplot(Caracteristiques_membres_menage, aes(x=1,y=AGFEXACT_01_Q19_1)) + g
eom_boxplot()
```



Export d'un graphique

R permet de sauvegarder les graphiques sous plusieurs formats. Pour cela, il suffit de cliquer sur le graphique (bouton gauche) puis d'aller dans Fichier ! Sauver sous. Il est également

possible de faire directement un Copier-Coller vers un autre document. Pour cela, cliquez sur le graphique (bouton droit) puis sélectionnez Copier comme bitmap. Vous pouvez ensuite faire un Coller sous Open Office ou sous Word.

1. Principe de l'analyse bivariée

L'analyse bivariée consiste à étudier deux variables conjointement, puis éventuellement à tester le lien entre les deux variables.

Variables	Test paramétrique	Diagnostic	Test non paramétrique
Qualitative & Qualitative	χ^2	1. Les valeurs de toutes les cases du tableau des effectifs attendus doivent être supérieures ou égales a 5.	Test exact de Fisher
Qualitative (2 classes) & Numérique	T de Student	1. Les écart types sont égaux 2. Pour chaque groupe, la variable numérique suit une loi normale OU les effectifs sont supérieurs à 30.	Test des rangs de Wilcoxon
Qualitative (3 classes et plus) & Numérique	F de Fisher (ANOVA)	1. Les écart types sont égaux 2. Pour chaque groupe, la variable numérique suit une loi normale OU les effectifs sont supérieurs à 30.	Test de Kruskal-Wallis
Numérique & Numérique	R de Pearson	1. Au moins une des deux variables suit une loi normale.	R de Spearman

2. Analyse bivariée

Effectifs, centralité et dispersion

Les effectifs s'obtiennent avec l'instruction table à laquelle on doit maintenant fournir les deux variables au lieu d'une seule. Comme pour l'analyse univariée, parler d'effectif n'a pas vraiment de sens avec les variables continues ; seules les nominales, ordonnées et discrètes sont concernées.

```
>table(Caracteristiques_membres_menage$SEXF_01_Q18,Caracteristiques_membres_menage$AGFEXACT_01_Q19_1)
```

```
     0  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
24 25 26 27 28
  1  2 23 33 25 37 43 41 28 40 28 40 22 36 29 38 41 36 31 43 35 44 30 39 26
36 47 30 34 32
  2  3 28 29 36 25 29 34 19 28 23 34 28 36 36 35 42 31 29 28 27 47 25 43 29
33 33 48 33 35

    29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
53 54 55 56 57
  1 21 40 13 38 20 24 23 32 18 27 24 50 19 28 23 19 41 22 17 21 10 45 13 25
22 34 25 28 14
  2 16 54 20 30 21 21 37 35 22 33 20 66 14 31 17 17 39 17 10 42 17 61  7 20
14 15 22 18  5

    58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81
82 84 85 86 88
  1 13 14 43  9 23  9  8 19  6  5  4  3 16  4  4  1  3  4  3  4  1  2  7  1
1  3  1  0  1
  2 12 10 32  4  9  3  2  8  7  3  1  1 15  3  3  2  3  9  2  0  1  1 15  0
0  1  1  1  0

    89 95 100 110
  1  2  1   0   0
  2  0  0   1   2
```

Les indices de centralité et dispersion n'existent pas en bivariée. Par contre, il est possible de les calculer relativement à une autre variable.

La moyenne des âges pour les hommes :

```
> mean ( Caracteristiques_membres_menage$AGFEXACT_01_Q19_1[Caracteristiques
_membres_menage$SEXF_01_Q18==1],na.rm=TRUE)
[1] 31.14479
```

La moyenne des âges pour les femmes :

```
> mean ( Caracteristiques_membres_menage$AGFEXACT_01_Q19_1[Caracteristiques
_membres_menage$SEXF_01_Q18==2],na.rm=TRUE)
[1] 30.44636
```

L'écart type des hommes :

```
> sd( Caracteristiques_membres_menage$AGFEXACT_01_Q19_1[Caracteristiques_me
mbres_menage$SEXF_01_Q18==1],na.rm=TRUE)
[1] 19.62483
```

L'écart type des femmes :

```
> sd( Caracteristiques_membres_menage$AGFEXACT_01_Q19_1[Caracteristiques_membres_menage$SEXF_01_Q18==2],na.rm=TRUE)
[1] 18.74037
```

3. Représentation graphique bivariée

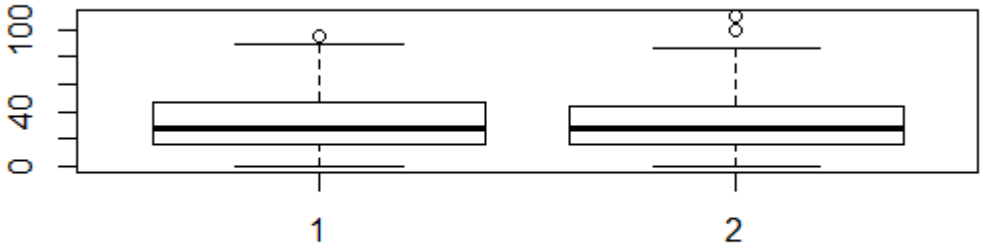
Deux qualitatives

Il n'existe pas vraiment de représentation graphique canonique pour deux variables qualitatives.

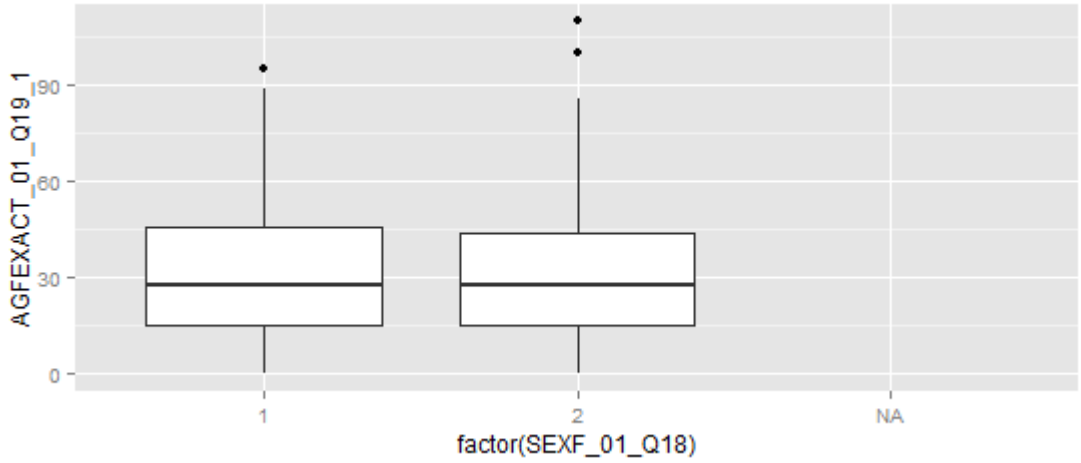
Qualitative & numérique

Pour une qualitative et une numérique, il est intéressant de graphiquement représenter des boites à moustache côte à côte, une pour chaque modalité de la qualitative :

```
>boxplot(Caracteristiques_membres_menage$AGFEXACT_01_Q19_1 ~ Caracteristiques_membres_menage$SEXF_01_Q18)
```



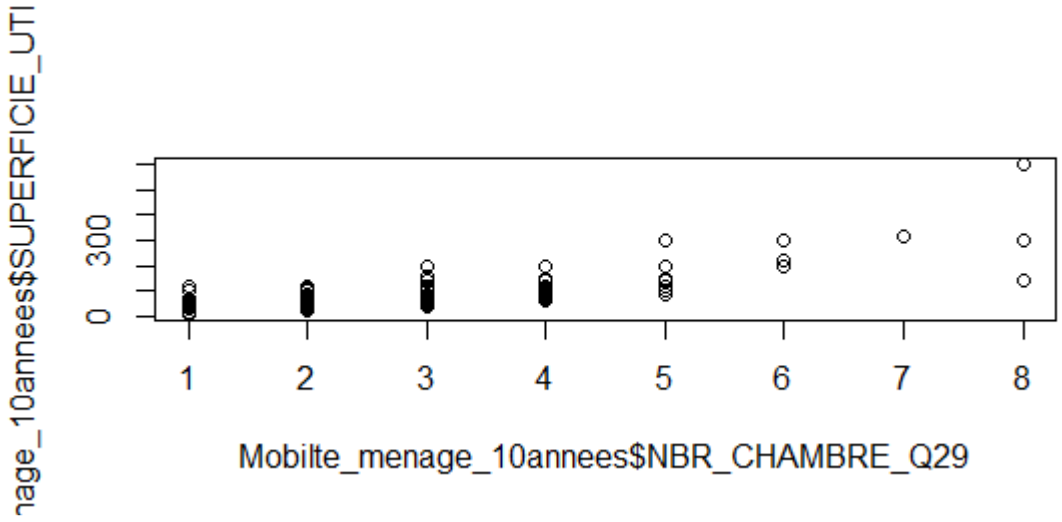

```
>ggplot(Caracteristiques_membres_menage, aes(y=AGFEXACT_01_Q19_1,x=factor(SEXF_01_Q18))) + geom_boxplot()
```



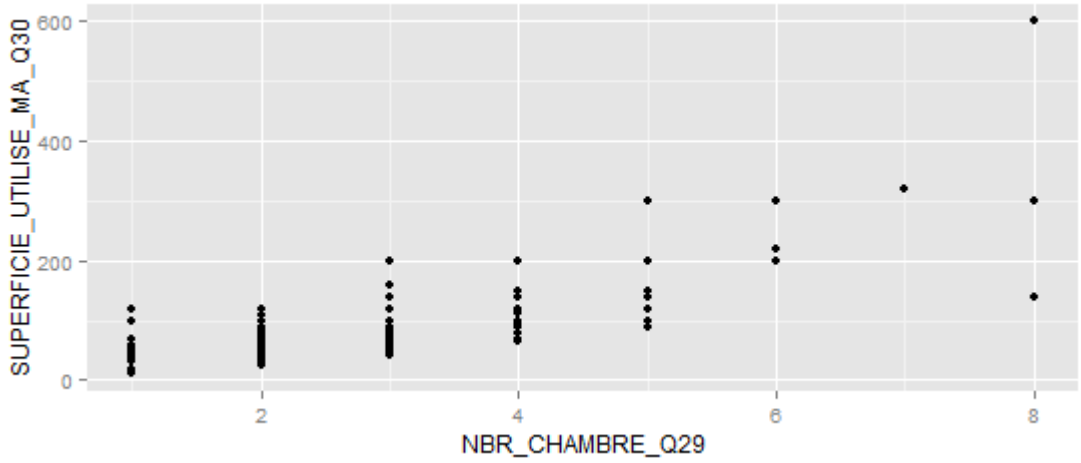
Deux numériques

Pour deux numériques, on peut tracer un nuage de points :

```
> plot(Mobilte_menage_10annees$NBR_CHAMBRE_Q29,Mobilte_menage_10annees$SUPERFICIE_UTILISE_MA_Q30)
```




```
> ggplot(Mobilte_menage_10annees, aes(x=NBR_CHAMBRE_Q29,y=SUPERFICIE_UTILISE_MA_Q30)) + geom_point()
```



4. Tests

Tester, c'est répondre à la question : y a-t-il un lien entre mes deux variables ?. Pour répondre à cette question, il existe deux types de tests. Les tests paramétriques sont des tests puissants mais ils nécessitent que les variables aient certaines propriétés. Les tests non-paramétriques sont moins puissants, mais n'imposent pas de condition d'application.

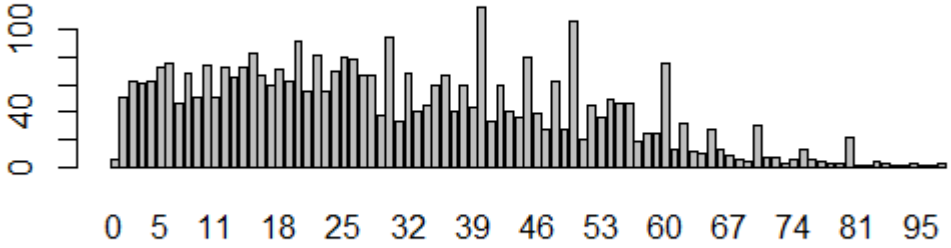
Le choix d'un test se fait donc en deux étapes :

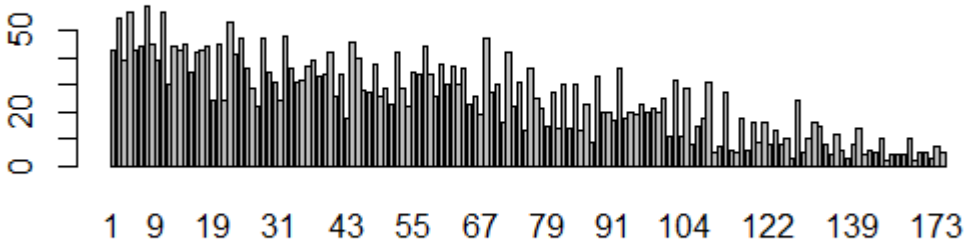
1. Le type des variables restreint le choix à un test paramétrique ou un test non paramétrique ;
2. Les propriétés des variables permettent de choisir entre le paramétrique et le non paramétrique.

Numérique & Numérique

Les tests possibles sont la corrélation de Pearson (paramétrique) et la corrélation de Spearman (non paramétrique). La condition d'application est : Au moins une des deux variables doit suivre une loi normale.

```
> barplot(table(Caracteristiques_membres_menage$AGFEXACT_01_Q19_1))
> barplot(table(Caracteristiques_membres_menage$NUM_LOGEM_DISTRICT_Q6))
```





Ne suivent pas la loi normale !

```
> cor.test(Caracteristiques_membres_menage$SEXF_01_Q18, Caracteristiques_me
mbres_menage$NUM_LOGEM_DISTRICT_Q6, method ="spearman")

        Spearman's rank correlation rho

data:  Caracteristiques_membres_menage$SEXF_01_Q18 and Caracteristiques_mem
bres_menage$NUM_LOGEM_DISTRICT_Q6
S = 8770900000, p-value = 0.2679
alternative hypothesis: true rho is not equal to 0
sample estimates:
        rho
-0.01815731

Warning message:
In cor.test.default(Caracteristiques_membres_menage$SEXF_01_Q18,  :
  Impossible de calculer la p-value exacte avec des ex-aequos
```

Le p-value étant petit -> pas de corrélation.

Qualitative & Qualitative

Pour deux variables qualitatives, le test à utiliser est le test du X^2 (paramétrique) ou le test exact de Fisher (non paramétrique). La condition nécessaire pour pouvoir utiliser le test du X^2 est la suivante :

1. les valeurs de toutes les cases du tableau des effectifs attendus doivent être supérieures à 5.

Remarque : Voir d'autres exemples de tests dans le help.