Agence Nationale de Statistique et de la Démographie

ANSD

Ecole Nationale de la Statistique et de l'Analyse Economique

ENSAE Pierre NDIAYE

Apurement des bases

Rédigé par :
Elisée AMEWOUAME et Brahima TOU
Elèves ingénieurs statisticiens
économistes

Sous la supervision de :
Mouhamadou Hady DIALLO
Ingénieur statisticien, Data Scientist

28 Juin 2022

Contents

I- Apurement des bases

Dans cette partie, nous allons apurer la base en vue de la préparer aux divers analyses.Elle consistera principalement à visualiser la baser,à renommer certaines variables pour qu'elles soient comprehensibles,à detecter les valeurs manquantes et les imputer si neccessaires.

1- Section 14a

Nous importons la base section14a qui traite entre autres de l'impact socio-économique du covid-19 sur les ménages.

```
## # A tibble: 6 x 22
##   interview__id    grappe id_menage vague s14aq00 interview__key s14aq01 s14aq02
##   <chr>             <dbl>     <dbl> <dbl>   <dbl> <chr>          <dbl+l> <dbl+l>
## 1 1df091103eb748a~    343        11     1       5 00-01-63-80    9 [Ren~ 2 [Non]
## 2 1df091103eb748a~    343        11     1       5 00-01-63-80    8 [Ren~ 2 [Non]
## 3 1df091103eb748a~    343        11     1       5 00-01-63-80    4 [Sub~ 2 [Non]
## 4 1df091103eb748a~    343        11     1       5 00-01-63-80    2 [Réd~ 1 [Oui]
## 5 1df091103eb748a~    343        11     1       5 00-01-63-80    1 [Eté~ 2 [Non]
## 6 1df091103eb748a~    343        11     1       5 00-01-63-80    6 [Ren~ 2 [Non]
## # ... with 14 more variables: s14aq02a <dbl>, s14aq08 <dbl+lbl>, s14aq09 <dbl>,
## #   s14aq02b <dbl>, s14aq03 <dbl+lbl>, s14aq04 <dbl>, s14aq05 <dbl+lbl>,
## #   s14aq06 <dbl>, s14aq07 <dbl+lbl>, s00q08 <dbl+lbl>, s00q09 <dbl+lbl>,
## #   s00q28 <dbl+lbl>, s00q27 <dbl+lbl>, s00q00obs <chr>
```

```
## [1] 23523    22
```

La base section14a contient 23523 observations pour 22 variables que nous allons renommer avant d'en choisir les plus utiles.

Renommons les variables

```
## [1] TRUE
```

```
##  [1] "interview__id"      "grappe"             "id_menage"
##  [4] "vague"              "repondant"          "interview__key"
##  [7] "event"              "event_occur"        "nbr_pers"
## [10] "solution"           "event_time"         "s14aq02b"
## [13] "menage_member"      "num_ordre"          "sexe"
## [16] "age"                "relationship"       "resultat_interview"
## [19] "motif"              "langue"             "quiz_result"
## [22] "obs"
```

Ainsi,nous avons renommer toutes les variables de la base afin de faciliter les comprehension des variables. L'objectif étant de déterminer les impacts du Covid-19 sur les ménages, nous allons nous limiter aux variables clés qui nous conduiront dans cette analyse.

Nous allons donc choisir les variables 'interview___id' pour l'identité du ménage, 'event' pour le type d'impact, 'event_occur' pour vérifier si l'impact a été observée ou non, 'nbr_pers' pour connaitre le nbre de personnes ayant subi l'impact, 'num_ordre' pour identifier l'individu du ménage qui a subi l'impact, ...

Selection des variables

```
##  [1] "interview__id" "event"          "event_occur"    "nbr_pers"
##  [5] "num_ordre"     "menage_member" "sexe"           "age"
##  [9] "relationship"  "solution"       "event_time"
```

Labelisation des variables

```
## # A tibble: 6 x 11
##   interview__id   event event_occur nbr_pers num_ordre menage_member sexe    age
##   <chr>           <fct> <fct>          <dbl>     <dbl> <fct>         <fct> <dbl>
## 1 1df091103eb748~ Reno~ Non               NA        NA <NA>          <NA>     NA
## 2 1df091103eb748~ Reno~ Non               NA        NA <NA>          <NA>     NA
## 3 1df091103eb748~ Subi~ Non               NA        NA <NA>          <NA>     NA
## 4 1df091103eb748~ Rédu~ Oui                1         2 Oui           <NA>     NA
## 5 1df091103eb748~ Eté ~ Non               NA        NA <NA>          <NA>     NA
```

```
## 6 1df091103eb748~ Reno~ Non                NA        NA <NA>          <NA>     NA
## # ... with 3 more variables: relationship <fct>, solution <fct>,
## #   event_time <dbl>
```

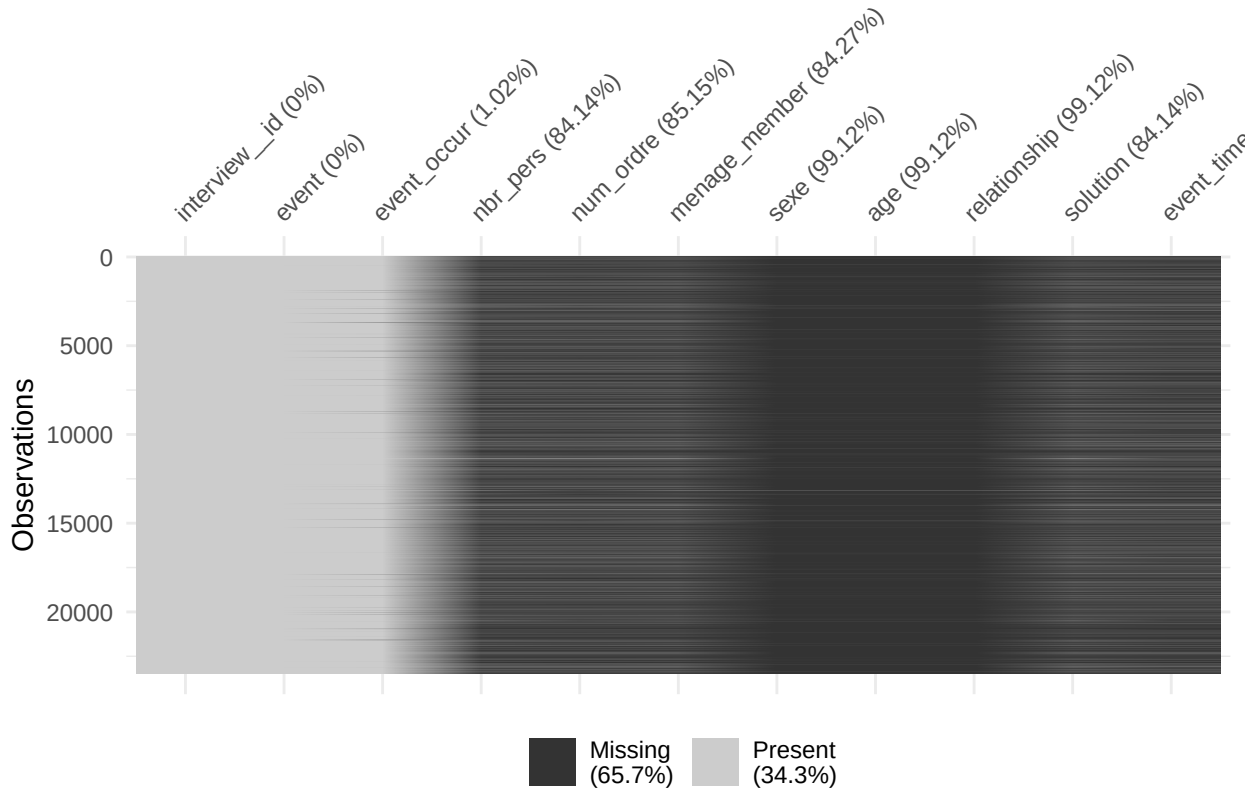
A présent que nous avons renommer et labéliser les variables dont nous auront besoin, nous allons supprimer les doublons contenus dans la base. Il est important de noter qu'il est impossible d'avoir deux lignes identiques sans que ce ne soit une erreur. En effet, nous avons retenu la variable 'num_ordre' qui désigne le numéro d'ordre de l'individu ayant subi l'impact (la variable 'event') et permet donc de distinguer les doublons.

Supression des doublons

```
## [1] 61
```

Nous avons supprimer 61 doublons. A présent, nous allons visualiser les valeurs manquantes de notre base de données.

Valeurs manquantes et imputations

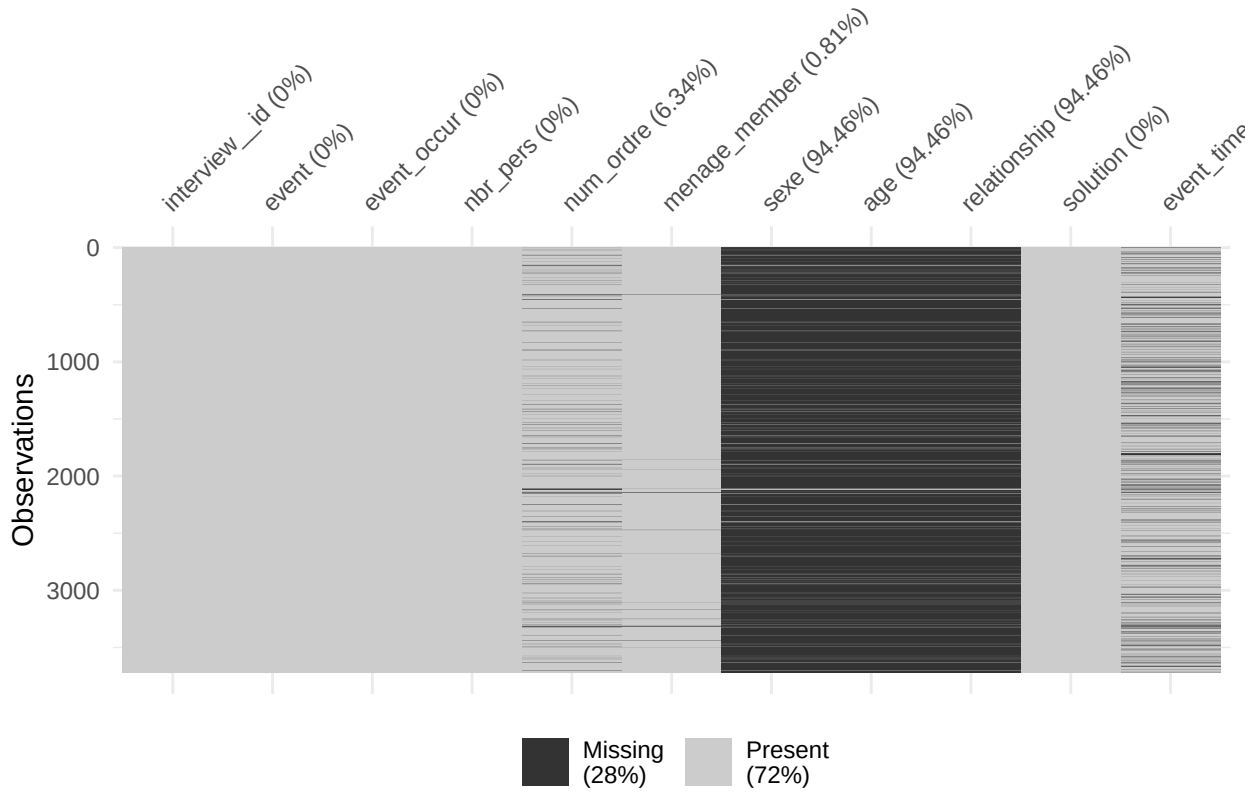


les variables les plus importantes comportent presque entièrement des valeurs manquantes (au delà de 80%). On ne peut donc pas les imputer ou les supprimer sans courir le risque

de perdre de l'information. Cependant, dans notre analyse, nous nous interesserons beaucoup plus à ceux qui ont ont subi les impacts (ceux qui ont dit "oui" pour la variable "event_occur"). Mais avant cela, nous allons éliminer les valeurs manquantes de la variable 'event_occur' car elle n'en contient que 1% et il est compliqué de les imputer car de cette variable dépend plusieurs autres.

```
## [1] FALSE
```

Maintenant que les valeurs manquantes de la variable 'event_occur' sont éliminées, nous allons filtrer les données selon ceux qui ont subi un impact et visualiser de nouveau les valeurs manquantes.



- les variables "age", "sexe" et "relationship" contiennent chacune plus de 90% de valeurs manquantes donc difficiles à exploiter. Nous allons donc les supprimer de la base.

```
## [1] "interview__id" "event"         "event_occur"   "nbr_pers"
## [5] "num_ordre"     "menage_member" "solution"      "event_time"
```

Mis à part ces 3 variables, les autres variables comme "solution", "event_time", "event_occur", "nbr_pers"… contient peu ou pas de valeurs manquantes.

- Certaines valeurs manquantes sont explicables car cela dépend d'autres variables. Par exemple, si aucune solution n'a été trouvée, on ne peut pas donner le temps que l'impact a pris avant de disparaitre. Il y aura donc nécessairement une valeur manquante dans la variable "event_time". Nous allons donc traiter ces données en fonction de nos objectifs.

- La variable "num_order" fournit juste le numéro d'ordre de l'individu impacté. Elle permet donc de détecter les doublons au cas où deux personnes dans le même ménage auraient eu le même problème. Cependant, on peut pas imputer un numéro d'ordre car ce numéro est unique pour chaque individu. Enfin, la variable "menage_member" qui détermine si l'individu est toujours dans le ménage ne contient que très peu de valeurs manquantes (0,81%).

2- Section 14b

Importation de la section 14b

```
## # A tibble: 6 x 48
##   interview__id  grappe id_menage vague s14bq00 interview__key   s14bq01 s14bq02
##   <chr>           <dbl>     <dbl> <dbl>   <dbl> <chr>          <dbl+lbl> <dbl+l>
## 1 1df091103eb74~    343        11     1       5 00-01-63-80    111 [Pri~ 2 [Non]
## 2 1df091103eb74~    343        11     1       5 00-01-63-80    120 [Att~ 2 [Non]
## 3 1df091103eb74~    343        11     1       5 00-01-63-80    102 [Déc~ 2 [Non]
## 4 1df091103eb74~    343        11     1       5 00-01-63-80    105 [Ino~ 2 [Non]
## 5 1df091103eb74~    343        11     1       5 00-01-63-80    113 [Per~ 2 [Non]
## 6 1df091103eb74~    343        11     1       5 00-01-63-80    114 [Fai~ 2 [Non]
## # ... with 40 more variables: s14bq03a <dbl+lbl>, s14bq03b <dbl>,
## #   s14bq04a <dbl+lbl>, s14bq04b <dbl+lbl>, s14bq04c <dbl+lbl>,
## #   s14bq04d <dbl+lbl>, s14bq04e <dbl+lbl>, s14bq04f <dbl+lbl>,
## #   s14bq05__1 <dbl>, s14bq05__2 <dbl>, s14bq05__3 <dbl>, s14bq05__4 <dbl>,
## #   s14bq05__5 <dbl>, s14bq05__6 <dbl>, s14bq05__7 <dbl>, s14bq05__8 <dbl>,
## #   s14bq05__9 <dbl>, s14bq05__10 <dbl>, s14bq05__11 <dbl>, s14bq05__12 <dbl>,
## #   s14bq05__13 <dbl>, s14bq05__14 <dbl>, s14bq05__15 <dbl>, ...
```

```
## [1] 54670    48
```

La base comporte 54670 lignes et 48 variables.Dans la suite, nous renommerons les variables et les traiter.

Renommons les variables

Tout comme la base précedente nous donnerons des noms à certaines variables si neccessaires pour faciliter les analyses et commentaires.

```
## [1] TRUE
```

La vérification étant faite, on remplace les anciennes colonnes

```
##  [1] "interview__id"          "grappe"                 "id_menage"
##  [4] "vague"                  "repondant"              "interview__key"
##  [7] "event"                  "event_occur"            "date"
## [10] "annee"                  "revenu"                 "avoirs"
## [13] "prod_agri"              "cheptel"                "stock_aliment"
## [16] "achat_aliment"          "epargne"                "aide_parent"
## [19] "aide_gouv"              "aide_ONG"               "marier_une_fille"
## [22] "change_habit"           "cheaper_food"           "extra_job"
## [25] "emploi_chomeur"         "emploi_enfant"          "descolarisation_enfant"
## [28] "migration"              "reduction_depenses"     "credit"
## [31] "vente_actif_agric"      "vente_bien_durable"     "vente_terrain"
## [34] "louer"                  "vente_vivres"           "peche"
## [37] "ente_betail"            "confiage_enfant"        "activite_spirituel"
## [40] "culture_contre_saison"  "autre_statégie"         "aucune_strategie"
## [43] "strategie"              "resultat_interview"     "motif"
## [46] "langue"                 "quiz_result"            "obs"
```

Nous allons à présent sélectionner les variables qui seront les plus utiles. Nous avons plusieurs variables qui vont servir. Alors, plutot que de sélectionner, nous allons éliminer celles qui ne nous serviront pas. Il s'agira des variables 'grappe', 'vague', 'repondant', 'interview___key', 'resultat_interview', 'motif', 'langue', 'quiz_result', 'obs'.

```
##  [1] "interview__id"          "event"                  "event_occur"
##  [4] "date"                   "annee"                  "revenu"
##  [7] "avoirs"                 "prod_agri"              "cheptel"
## [10] "stock_aliment"          "achat_aliment"          "epargne"
## [13] "aide_parent"            "aide_gouv"              "aide_ONG"
## [16] "marier_une_fille"       "change_habit"           "cheaper_food"
## [19] "extra_job"              "emploi_chomeur"         "emploi_enfant"
## [22] "descolarisation_enfant" "migration"              "reduction_depenses"
## [25] "credit"                 "vente_actif_agric"      "vente_bien_durable"
## [28] "vente_terrain"          "louer"                  "vente_vivres"
## [31] "peche"                  "ente_betail"            "confiage_enfant"
## [34] "activite_spirituel"     "culture_contre_saison"  "autre_statégie"
## [37] "aucune_strategie"       "strategie"
```

Nous vérifions la présence ou non des doublons.

```
## [1] 0
```

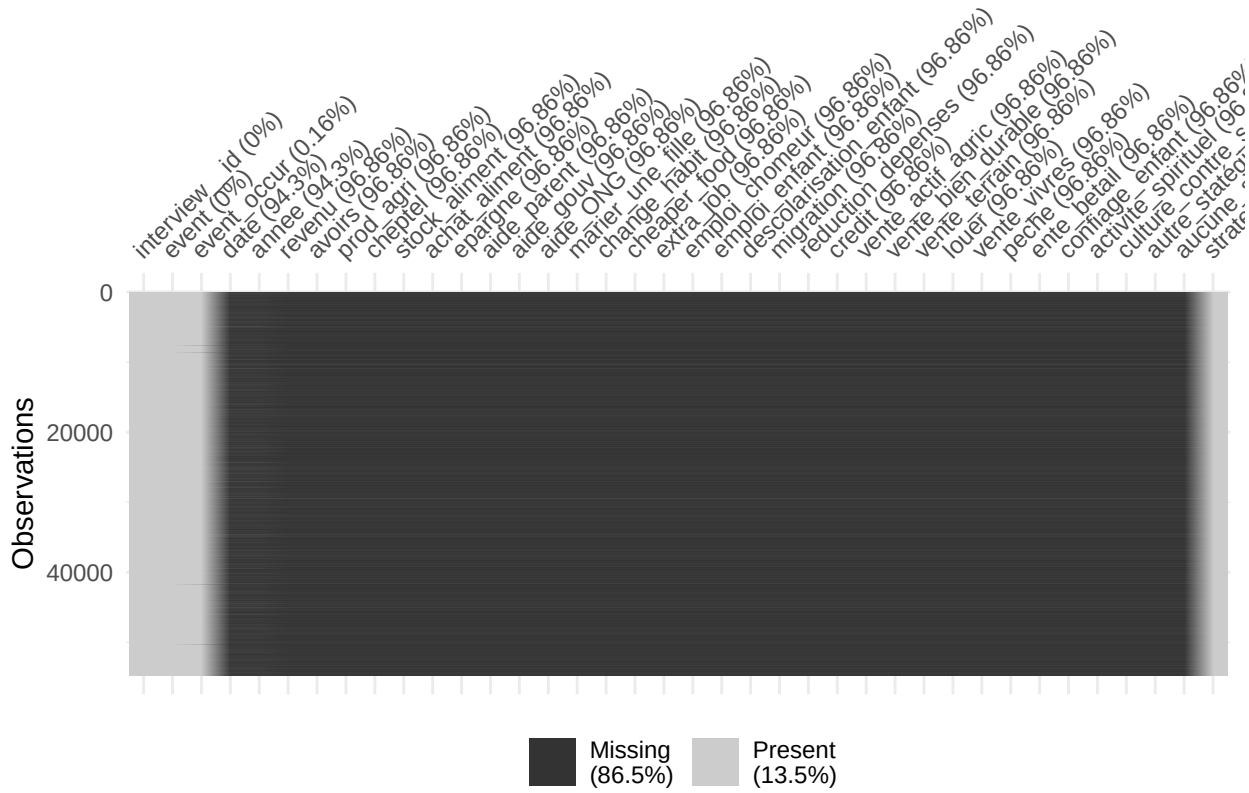
Nous notons donc la présence de 0 doublons.

Labellisation les variables

```
## # A tibble: 6 x 38
##   interview__id    event event_occur date  annee revenu avoirs prod_agri cheptel
##   <chr>            <fct> <fct>       <fct> <dbl> <fct>  <fct>  <fct>     <fct>
## 1 1df091103eb748a~ Prix~ Non         <NA>     NA <NA>   <NA>   <NA>      <NA>
## 2 1df091103eb748a~ Atta~ Non         <NA>     NA <NA>   <NA>   <NA>      <NA>
## 3 1df091103eb748a~ Décè~ Non         <NA>     NA <NA>   <NA>   <NA>      <NA>
## 4 1df091103eb748a~ Inon~ Non         <NA>     NA <NA>   <NA>   <NA>      <NA>
## 5 1df091103eb748a~ Pert~ Non         <NA>     NA <NA>   <NA>   <NA>      <NA>
## 6 1df091103eb748a~ Fail~ Non         <NA>     NA <NA>   <NA>   <NA>      <NA>
## # ... with 29 more variables: stock_aliment <fct>, achat_aliment <fct>,
## #   epargne <dbl>, aide_parent <dbl>, aide_gouv <dbl>, aide_ONG <dbl>,
## #   marier_une_fille <dbl>, change_habit <dbl>, cheaper_food <dbl>,
## #   extra_job <dbl>, emploi_chomeur <dbl>, emploi_enfant <dbl>,
## #   descolarisation_enfant <dbl>, migration <dbl>, reduction_depenses <dbl>,
## #   credit <dbl>, vente_actif_agric <dbl>, vente_bien_durable <dbl>,
## #   vente_terrain <dbl>, louer <dbl>, vente_vivres <dbl>, peche <dbl>, ...
```

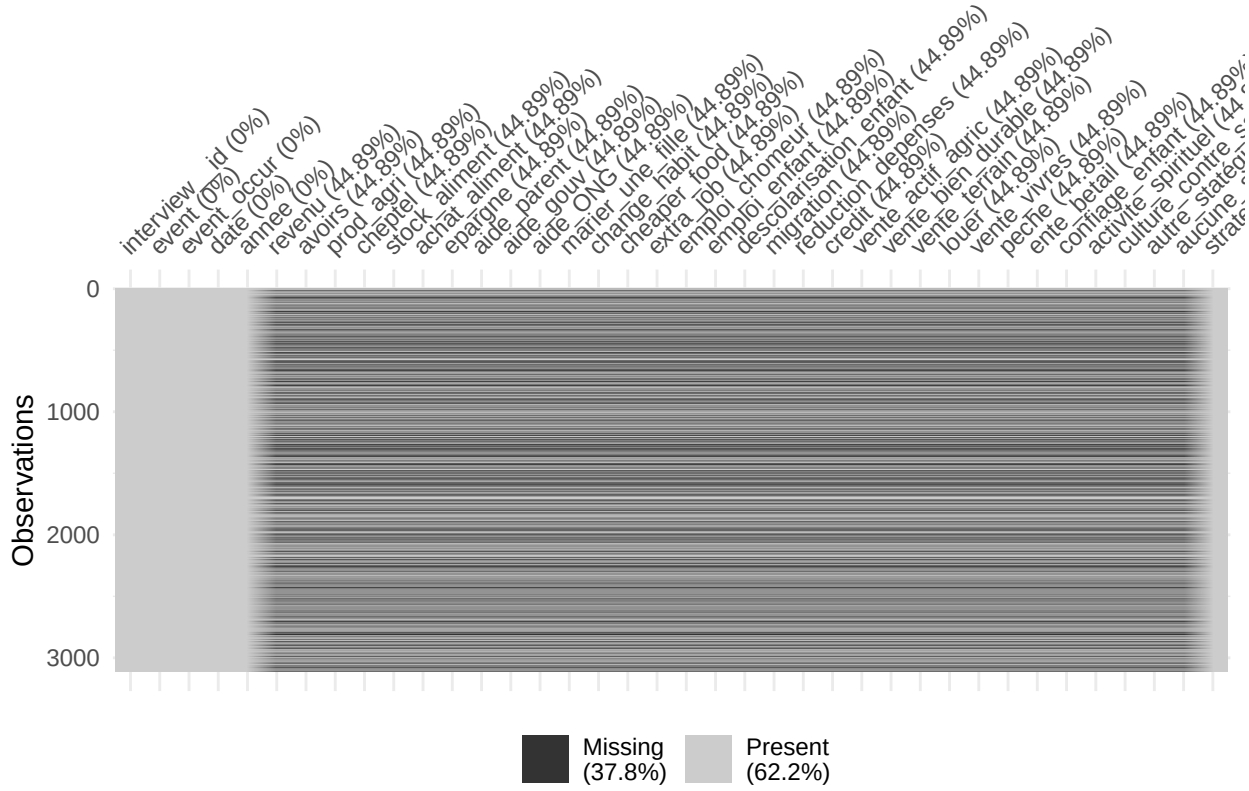
Nous nous intéressons à présent au valeurs manquantes

Valeurs manquantes et leur traitement



La base contient 86,5% de valeurs manquantes mais qu'on peut expliquer par le fait que certaines variables dépendent d'autres variables-clé comme pour la section 14a. L'une de ces variables-clé est la variable 'event_occur' qui détermine s'il y a eu survenue d'un impact ou non. Cette variable contient O,16% de valeurs manquantes, ce qui est négligeable. De plus, même si nous imputons cette variable, nous ne saurons compléter les autres variables qui en dépendent. Nous n'avons d'autres choix que de les éliminer.

```
## [1] FALSE
```



Nous disposons de moins de valeurs manquantes quand nous nous interessons uniquement à ceux qui ont subi un impact. Cependant, nous remarquons que 44,89% des valeurs des variables à partir de la variable 'revenu' sont des valeurs manquantes. Cela est normal. En effet, chaque ligne correspond à la réponse d'u ménage quant au fait qu'il a subi ou non un choc. Cependant, les conséquences des choccs et les stratégies utilisées ne dépendent pass du type de choc. Un même ménage utilise 3 sortes de stratégies indépendamment du type de choc et donc ces variables ne sont renseignées qu'une seule fois pour le même ménage. Le même ménage est itéré 22 fois sur les lignes car il y a 22 chocs. Cependant les réponses aux stratégies et aux conséquences des chocs ne sont itérées qu'une seule fois. C'est ce qui explique d'ailleurs qu'on ait les mêmes proportions de valeurs manquantes pour ces autres colonnes. Pour nos analyses, nous les supprimerons pour étudier les conséquences des chocs et les stratégies utilisées.