Analyse des fichiers

Contents

1 Programme

- Apprendre à utiliser R et créer un tableau de bord
- Appliquer des règles déterministes pour trouver des écarts à étudier
- Détection d'anomalies de manière simple avec une variable
- Détection d'anomalies avec des variables multiples (plusieurs méthodes voir illustration)
- Prédiction pour une variable d'intérêt

Le document est produit dans R Studio, et les résultats d'analyses sont affichés dans ce document. Mais le code R est caché.

2 Import des données

2.1 Fichiers mono

On peut créer une boucle pour importer les fichier efficacement. Je montre le code R ici.

```
mono19=fread("data/Mono/PM_ALM_MonoSupport_31122019.csv",skip=2)%>%
  mutate(PM_ALM=as.numeric(gsub(" ","",PM_ALM)))
mono20=fread("data/Mono/PM_ALM_MonoSupport_31122020.csv",skip=2)%>%
  mutate(PM_ALM=as.numeric(gsub(" ","",PM_ALM)))
mono21=fread("data/Mono/PM_ALM_MonoSupport_31122021.csv",skip=2)%>%
  mutate(PM_ALM=as.numeric(gsub(" ","",PM_ALM)))
```

On peut faire une jointure entre les différentes années, mais il semble que les numéros de polices n'ont pas été bien anonymisés. C'est à dire que le police 1000 de l'année 2019 ne correspond pas à la police 1000 de l'année 2020. donc on ne peut pas étudier l'évolution.

2.2 Fichiers multi

Problème de format a priori. Il faudrait refaire les extractions... Mais si c'est justement ça le problème, alors il est possible aussi de nettoyer les données pour avoir un format exploitable.

2.3 Rentes

On peut créer une boucle pour importer les fichier efficacement et effectuer quelques nettoyages.

3 Nettoyage simple des données

3.1 Mono

On peut afficher un résumé statistique des données. C'est une fonction de base dans R. Les résultats ne sont pas bien formatés. On peut faire un tableau de bord plus "joli".

```
##  CodeProduit           NoPolice      TypeDePolice       DateDeProduction   AnneeNaissanceAssure   Se
##  Length:120833      Min.   :     1   Length:120833      Length:120833      Min.   :1900         Min.
##  Class :character   1st Qu.: 30209   Class :character   Class :character   1st Qu.:1936         1st
##  Mode  :character   Median : 60417   Mode  :character   Mode  :character   Median :1954         Medi
##                     Mean   : 60417                                         Mean   :1955         Mean
##                     3rd Qu.: 90625                                         3rd Qu.:1970         3rd
##                     Max.   :120833                                         Max.   :2019         Max.
##      PM_ALM             TMG         AnneeFinGarantiePourTMG   Version
##  Min.   :      0   Min.   :0.0000   Min.   :2020            Length:120833
##  1st Qu.:      0   1st Qu.:0.0000   1st Qu.:2020            Class :character
##  Median :   1431   Median :0.2500   Median :2020            Mode  :character
##  Mean   :  33876   Mean   :0.9507   Mean   :2113
##  3rd Qu.:  17262   3rd Qu.:2.0000   3rd Qu.:2023
##  Max.   :6174645   Max.   :5.0000   Max.   :3000
```

On peut supprimer les variables qui sont manquantes partout.

```
## Warning in set(mono19, j = cols_to_delete, value = NULL): length(LHS)==0; no columns to delete or as
```

```
##  CodeProduit           NoPolice      TypeDePolice       DateDeProduction   AnneeNaissanceAssure   Se
##  Length:120833      Min.   :     1   Length:120833      Length:120833      Min.   :1900         Min.
##  Class :character   1st Qu.: 30209   Class :character   Class :character   1st Qu.:1936         1st
##  Mode  :character   Median : 60417   Mode  :character   Mode  :character   Median :1954         Medi
##                     Mean   : 60417                                         Mean   :1955         Mean
##                     3rd Qu.: 90625                                         3rd Qu.:1970         3rd
##                     Max.   :120833                                         Max.   :2019         Max.
##      PM_ALM             TMG         AnneeFinGarantiePourTMG   Version
##  Min.   :      0   Min.   :0.0000   Min.   :2020            Length:120833
##  1st Qu.:      0   1st Qu.:0.0000   1st Qu.:2020            Class :character
##  Median :   1431   Median :0.2500   Median :2020            Mode  :character
##  Mean   :  33876   Mean   :0.9507   Mean   :2113
##  3rd Qu.:  17262   3rd Qu.:2.0000   3rd Qu.:2023
##  Max.   :6174645   Max.   :5.0000   Max.   :3000
```

4 Analyses exploratoires

4.1 Mono

4.1.1 CodeProduit

On peut effectuer des analyses exploratoires classiques, et créer un tableau de bord de façon automatique. voir l'exemple html. Des fichiers html peuvent être rassembler pour créer un site internet. Et l'avantage des fichiers statiques, c'est qu'ils peuvent être déposés dans un dossier de serveur, accessible à tous par un navigateur web classique. En apparence, c'est un site web.

CodeProduit	N	PM_ALM
CEC	34621	3.318591e+09
CEF	20222	2.776546e+08
CEP	3277	1.407823e+08
CEVE	194	4.668597e+06
EEC	6394	1.591771e+08
VOE	3307	2.130334e+07
ECLOR	26912	3.845641e+05
EH	13559	1.285346e+08
ETR	307	1.937233e+06
OPA	7826	1.106908e+07
PEP	2238	1.288586e+07
PEV	1972	1.627347e+07
PRIMOR	4	6.297153e+04