# **Devoir 1- STA211- Prétraitement des données**

* **NOM** : Elio
* **PRENOM** : Bou Serhal
* **DATE** : 10 / 1 / 2023

Dans ce document on va voir comment traiter une base de donnes de zéro, autrement dit c’est comment travailler avec une base de données contenant des valeurs manquantes.

1. Tout d’abord on va importer cette donnée du site :

[https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data](https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data%20)

En utilisant la fonction « read.csv () », En changeant les noms des variables dans le but d’obtenir une data plus claire, alors on obtient le Data Frame réserve sur R.

1. Coup d'œil sur notre Data Frame pour voir la structure et le type :

Le type de cette Data est : une Data Frame

Le nombre d’observation est : 32560 observations de 15 variables

Les noms des variables : Age, Education\_Num, Occupation,….

Le type de chaque variable : Quantitative ou Qualitatives ou autres.

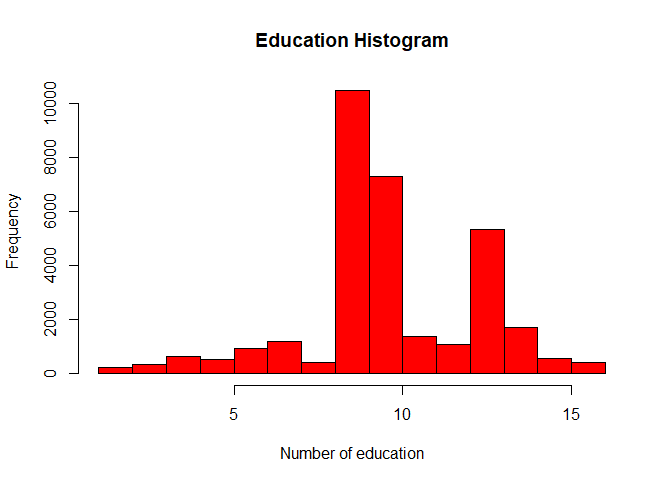
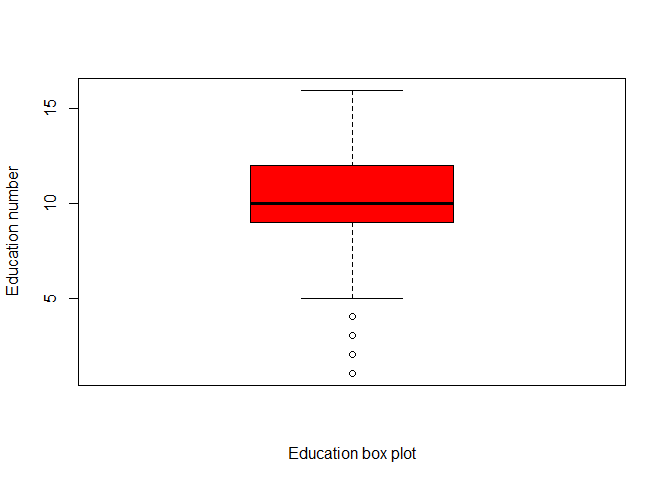
De plus on peut faire une petite compréhension des variables.

1. En utilisant la fonction « describe » de la librairie « Hmisc » on obtient la statistique descriptive

Alors pour la variables « Education\_Num » on peut voir le Max, Min, Médian, Moyenne…

1. Représentant Graphiquement la variables « Education\_Num » par :

Une boite a moustache et un Histogramme ou on peut visualiser la moyenne, les quantiles, la fréquence

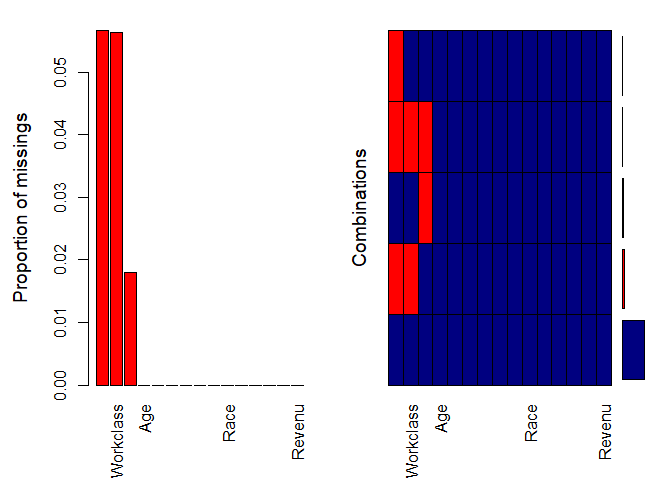


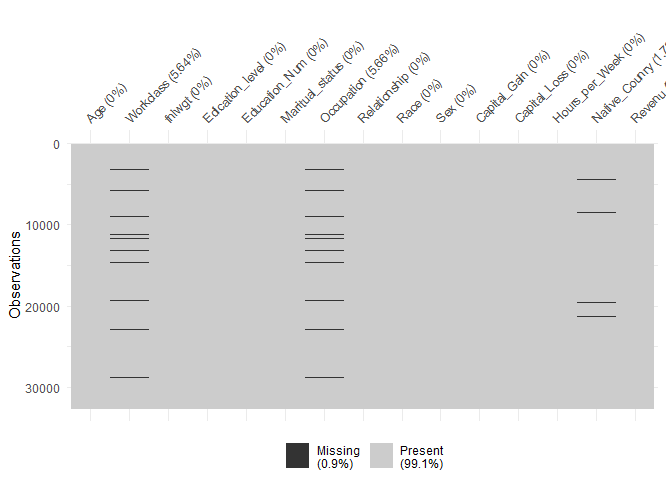
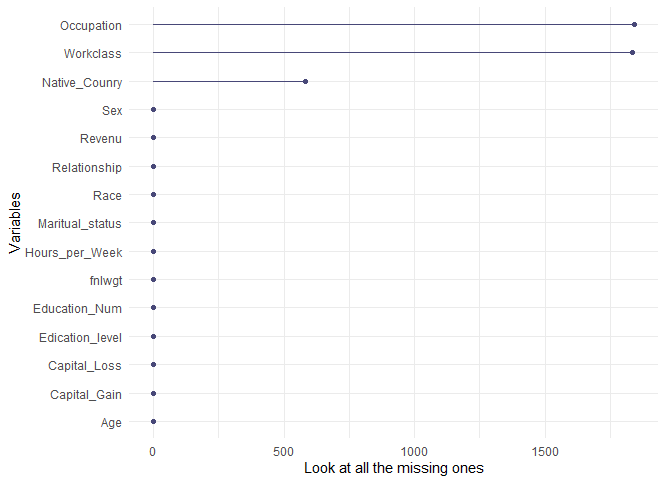
1. Notre donnée contient des valeurs manquantes sous forme d’une point d’interrogation « ? »

Alors que Rstudio ne considéré pas que c’est une valeur manquantes, c’est pour cela on va transformer ce symbole « ? » en un caractères « NA » : on définit **Data\_NA** la donne contenant les « NA »

Ensuite pour visualiser les variables contenant des NA :

On peut aussi savoir le % des NA des chaque ligne et chaque colonne.





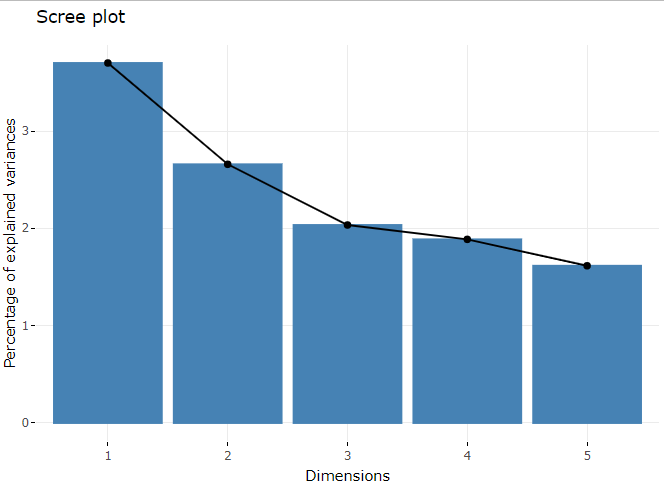
1. Maintenant éliminant les NA de note base de donnes.

On définit par « d1 » la nouvelle Data Frame qui ne contient pas des NA

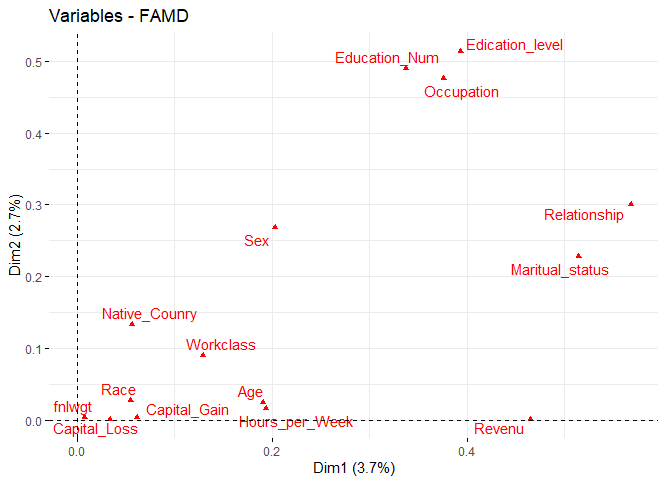
1. On constate qu’on peut réduire logiquement les modalités des variables : « éducation », « occupation », « workclass ».
2. Détectant les variables inutiles en utilisant l’analyse factorielle des donnes mixtes « AFDM »

On peut voir quelle variables contribuent ou qualités de représentations la plus élevés.

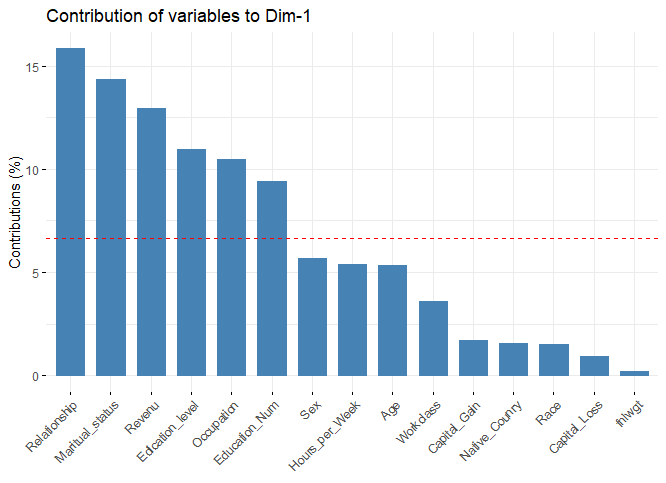
Tout d’abord on a pu réduit notre base de donne a 5 dimensions (5 valeurs propres), on peut les visualiser sur un histogramme, de plus on peut trouver les coordonnes de chaque variables sur les 5 dimensions, autrement dit : on peut trouver « Qualité de représentation » et « contribution » de chaque variables par rapport aux 5 dimensions (par rapport aux axes factorielles).

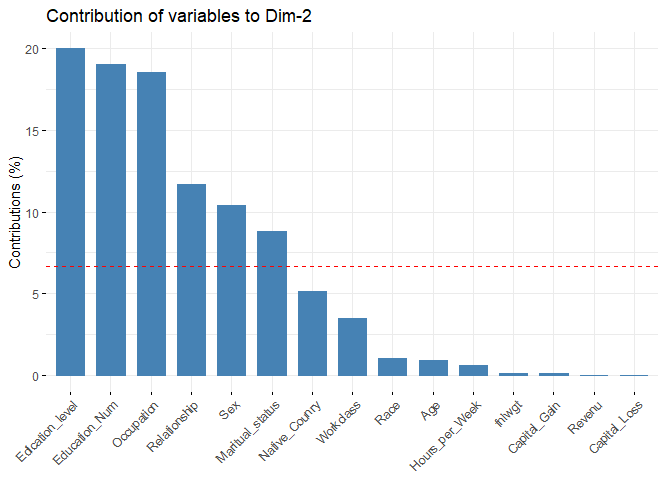


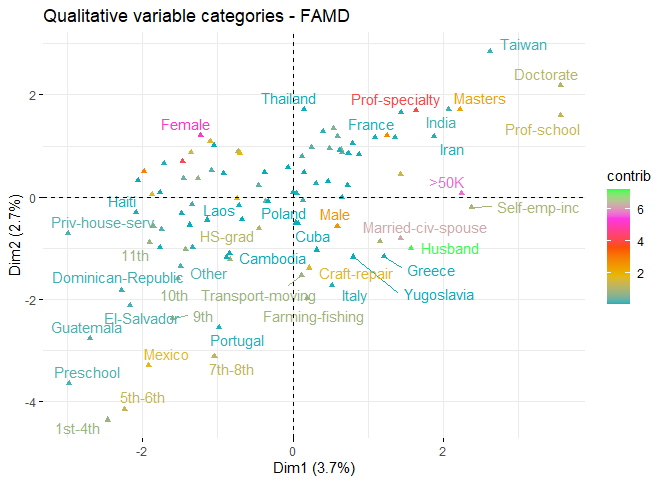
Maintenant on peut visualiser les contributions des variables par rapport aux axes factorielles, alors la variables la plus proche de l’axe a la contribution la plus élevés.

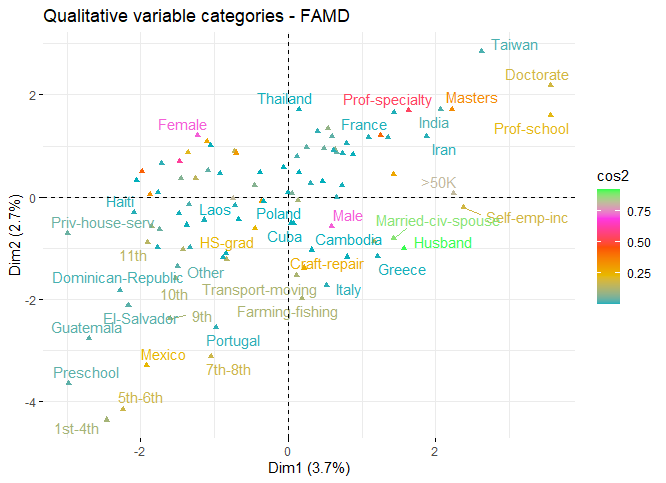


On peut voir la contribution de chaque variables par rapport à chaque axe sous forme d’un histogramme

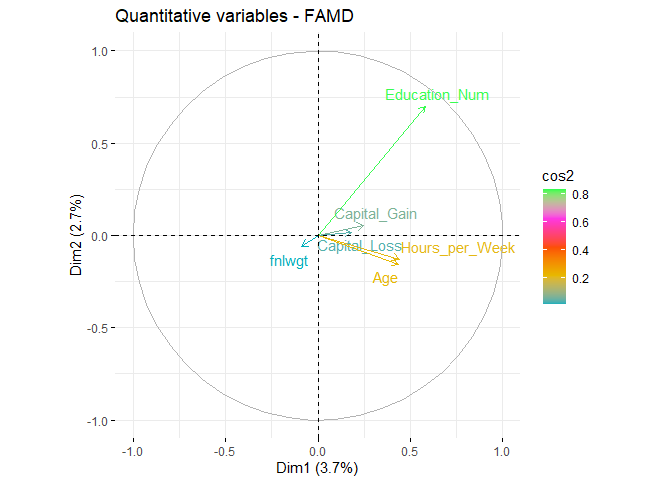


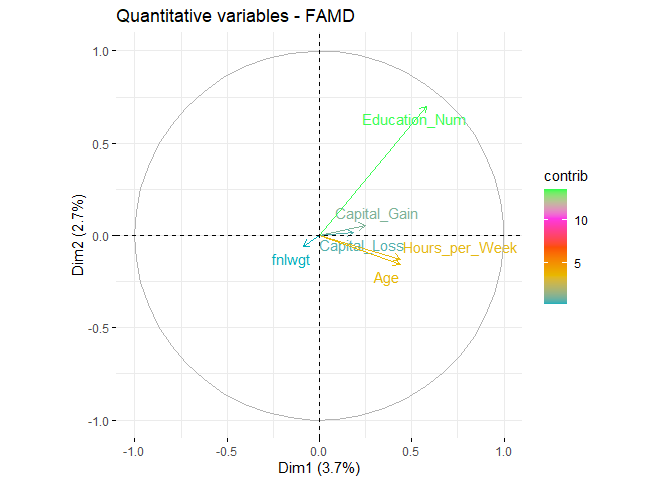


Maintenant, on va voir la « qualité de représentation » et «contribution » des variables QUALITATIVES :



Maintenant, on va voir la « qualité de représentation » et «contribution » des variables QUANTITATIVE :

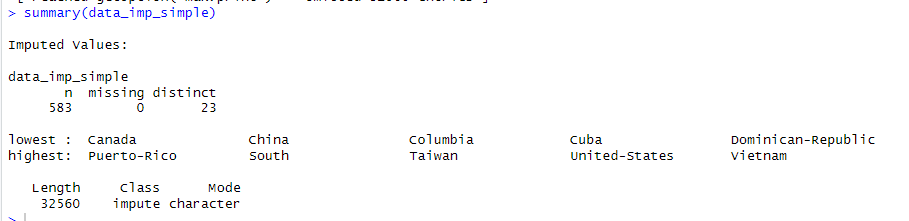




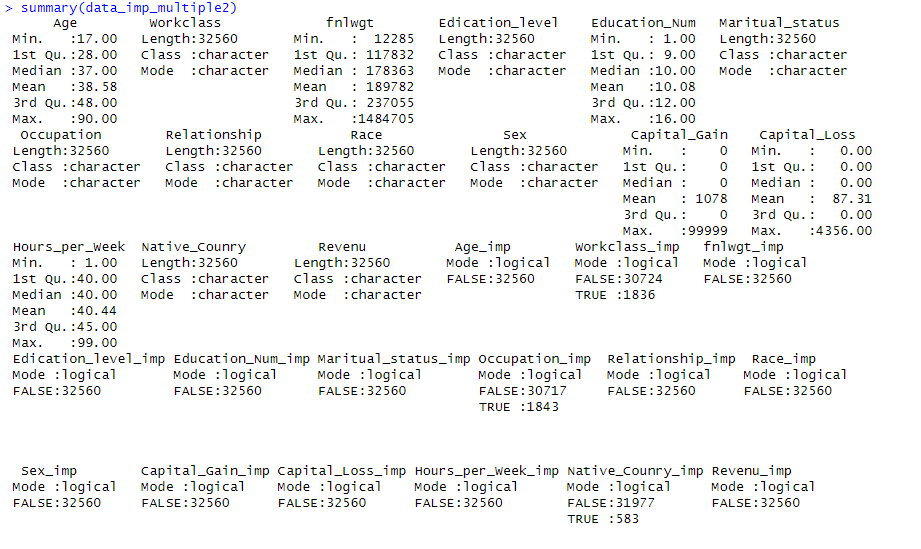
1. Pour imputer les valeurs manquantes « NA », il existe plusieurs méthodes d’imputations :

« Imputation simple » à laquelle on impute chaque NA (chaque variable), tandis que « imputation multiple » on impute les valeurs NA toutes ensembles mais sa dépend de combien de possibilités on a (chaque variable a un certain nombre de modalités).

Pour l’imputation simple :



Pour l’imputation multiple :



1. Par discrétisation on transforme les variables quantitatives en variables qualitatives :

Plusieurs methodes existes par exemple Methode ascendante de Chi Merge de Kerber, ou bien methode Descendante MDLP ou arbre de decision…

Mais la plus efficace entre eux c’est celle de MDKP (FAYYAD ET IRANI)

