# INFO308: Statistiques et Analyse des données

Année Académique 2020-2021

Dr. Paulin Melatagia

# Première séance de cours

1) Présentation de l’équipe pédagogique
2) Présentation du syllabus du cours
3) Chapitre 0 : Introduction

# Objectifs

# Statistiques et Analyse des données : objectifs du cours

1) Traiter et décrire l'information contenue dans des grands ensembles de données

2) Interpréter correctement les graphiques sur des données

3) Savoir dans quelle mesure les résultats obtenus sur un échantillon convenablement choisi apportent une connaissance fiable des caractéristiques de la population d'origine

# Plan du cours

# Statistiques et Analyse des données

1) Partie 1 : Analyse de données
    1) Observation des données unidimensionnelles
    2) Observation des données bidimensionnelles
    3) Séries temporelles
    4) Observation des données multi-dimensionnelles
2) Partie 2 : Statistique Inférentielle

# Statistiques et  Analyse des données

1) Partie 1 : Analyse de données
2) Partie 2 : Statistique Inférentielle
    1) Fondements
    2) Estimation
    3) Échantillonnage
    4) Introduction au Tests

# Pre-requis et Outils

# Statistiques et  Analyse des données

Prérequis

- Algèbre linéaire
- Statistique descriptive
- Théorie des probabilités

Outil

- Python

# Introduction

# Introduction à l’analyse de données

- L'analyse des données est une technique relativement ancienne 1930 (PEARSON, SPEARMAN, HOTELLING).
- Elle a connu cependant des développements récents 1960-1970 du fait de l'expansion de l'informatique.

# Introduction à l’analyse de données

- L'analyse des données est une technique d'analyse statistique d'ensemble de données.
- Elle cherche à décrire des tableaux et à en exhiber des relations pertinentes.
- Elle se distingue de l’analyse exploratoire des données.

# Introduction à l'analyse de données

- L'objectif de la démarche statistique est de faire apparaître ces liaisons.
- Les deux types de relations fondamentales sont les relations d'équivalence et les relations d'ordre. Ainsi, une population peut-elle être décomposée en classes hiérarchisées.

# Introduction à l’analyse de données

- Le but de l’analyse de données est de synthétiser, structurer l'information contenue dans des données multidimensionnelles (n individus, p variables)

# Introduction à l’analyse de données

- Le but de l’analyse de données est de synthétiser, structurer l'information contenue dans des données multidimensionnelles (n individus, p variables)
- L'analyse des données est utilisée dans tous les domaines dès lors que les données se présentent en trop grand nombre pour être appréhendées par l'esprit humain.

# Exploiter les 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 s 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 |

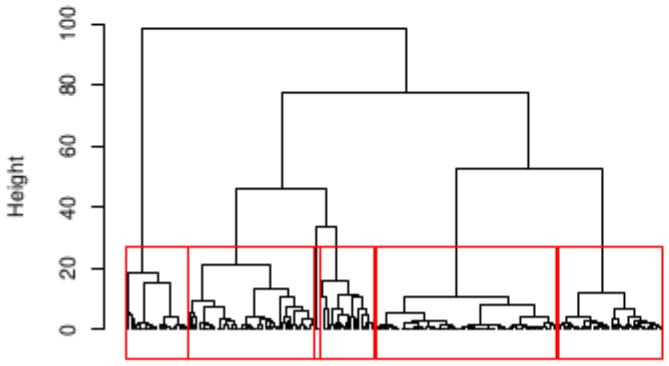
- Volume classique : milliers à millions de lignes, dizaine à centaines de colonnes
- Exploration systématique impossible (même pour de petits paquets de données)

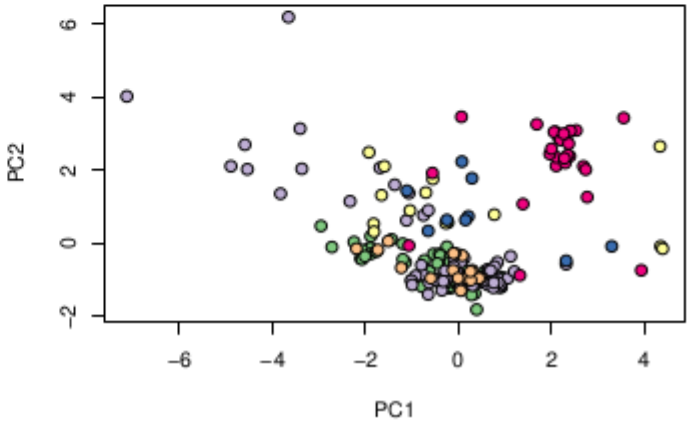
# Outils d'exploitation

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

# 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
- Stratégie :
  - analyse exploratoire
  - formulation d'hypothèses
  - construction de modèles pour valider les hypothèses

# Méthodes d’analyse de données

Deux groupes de méthodes :

1) méthodes de **classification**: réduire la taille de l’ensemble des individus en formant des groupes homogènes ;
2) Méthodes **factorielles**: réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques.

# Méthodes d'analyse de données