Statistiques descriptives

Le chat !

Objectifs du module

Transformer des données brutes en information

Ce module est le premier d'une série de 3 modules dédiés aux statistiques et aux probabilités :

1. **Statistiques descriptives**
   Topic = Electricity production and consumption in France
2. **Tests d'hypothèses**
   Topic : Détection de sons
3. **Linear regression**
   Topics: Hospital emergency department overcrowding, wine characteristics, housing price prediction, Moneyball!

La statistique descriptive est une branche des statistiques qui regroupe les nombreuses techniques utilisées pour décrire un ensemble relativement important de données. L'objectif étant d'organiser et de résumer cet ensemble afin de le rendre compréhensible et de mettre en lumière les informations qu'il contient.

Ce module est découpé en 3 parties :

1. Notions de base des statistiques descriptives
2. Application à un cas réel
3. Bonus :
   a. L'analyse de données multidimensionnelles avec l'ACP
   b. Étude de données particulières : les séries temporelles

Modalités

- Durée du projet : 4 jours
- Ce projet sera réalisé en autonomie
- On peut trouver beaucoup d'informations différentes dans le même set de données. N'hésitez pas à échanger avec vos voisins pour comparer les informations trouvées et les méthodes choisies pour les obtenir.
- Un jupyter notebook est fourni par étape. Chaque notebook contient :
  - Des rappels de cours : quelques notions importantes, des schémas ou des références de cours
  - Une liste « TODO » : des consignes et des questions ouvertes. Les questions posées sont là pour orienter l'exploration du dataset. Si cette exploration soulève d'autres questionnements, tentez d'y répondre : c'est l'objectif même de l'analyse exploratoire.
  - Une partie « aide » qui donne quelques conseils pour avancer

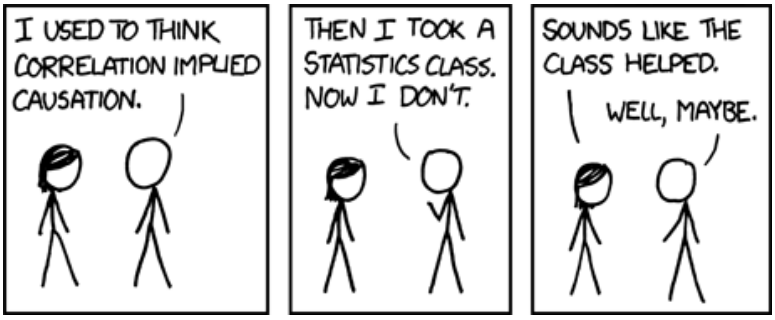
Étape 1

Analyse descriptive – cas théorique (0.5 jour)

Objectifs de l'activité

Cette partie a pour objectif de décrire et de comprendre les notions de base des statistiques univariées et multivariées. Les travaux seront effectués sur un dataset « artificiel », créé pour illustrer ces notions. Aucune compétence ne sera donc validée ici : une compétence ne peut être validée qu'après avoir été appliquée au cas d'étude réel de la partie suivante.

1. Vocabulaire
2. Mesures de tendance centrale, de dispersion, de forme
3. Relations entre variables
4. Notions de réduction de dimension



Compétences

- Il n'y a aucune compétence à valider dans cette partie

Consignes

- Utiliser le jupyter notebook 'Stat_descriptives_part1_découverte.ipynb', et suivre la trame proposée
- Si les différentes notions abordées ici vous sont déjà familières, vous pouvez passer vite à la partie suivante. A l'inverse, si ça n'est pas le cas, prenez autant de temps que nécessaire pour bien comprendre et assimiler les concepts présentés. Ils seront utilisés dans la partie suivante.

Étape 2

Analyse exploratoire d'un jeu de données réel (2,5 jours)

Objectifs de l'activité

L'objectif est ici d'appliquer les différentes notions que vous venez de voir à un jeu de données « réelles », représentant le paysage électrique français, afin d'en tirer des informations claires et fiables.

Vous allez vous rendre compte que les étapes de mise en forme, nettoyage et exploration de données peuvent être chronophage lorsqu'on traite de données réelles, qui n'ont pas été créées « par » ni « pour » des data scientists.

Elles n'en sont pas moins indispensables : en se lançant dans un problème de Machine Learning sans analyse exploratoire, on risque de perdre un temps précieux à explorer des voies qui n'en sont pas, à se questionner sur des comportements étranges, ou à passer à côté de voies prometteuses …

Les données que vous allez étudier dans cette partie sont donc les données « brutes », téléchargées directement depuis le site internet du RTE, et n'ayant subi aucun prétraitement. Ce set de données contient, pour plusieurs années :

- Consommations et sources de production régionales et nationales
- Émissions de C02
- Prix de l'électricité en France et en Allemagne
- Échanges aux frontières

1. Manipulation de données avec Pandas
   - Création des dataframes à partir des fichiers RTE
   - Exploration des données contenues dans ces dataframes

2. Étude des relations entre variables

Compétences

- Choix de descripteurs statistiques pour représenter une information
- Identification du lien entre différentes variables
- En bonus (sans validation) : manipulation de données avec Pandas !

Consignes

- Utiliser le jupyter notebook 'Stat_descriptives_part2_mise_en_pratique.ipynb', et suivre la trame proposée
- Répondre aux différentes questions en gardant en tête qu'elles ne constituent qu'une orientation à l'exploration du dataset
- Il est important ici d'échanger avec vos voisins pour comprendre à quel point des informations différentes peuvent être extraites du même jeu de données

Étape 3

Au choix une des deux activités : ACP ou séries temporelles (1 jour)

Focus sur l'ACP (1 jour)

Objectifs de l'activité

L'ACP (Analyse en Composantes Principales), appliquée à p variables quantitatives, a pour but de résumer les liens entre les variables par l'analyse des covariances ou des corrélations, et dresser une « carte » des individus indiquant leur position par rapport à ces liens.

Dans ce notebook, nous allons utiliser l'ACP à des fins exploratoires à l'aide du package Scikit Learn : https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Ouverture : l'ACP permet d'explorer mais pas que !

… Elle permet également de condenser l'information présente dans nos variables : on peut alors réduire notre dataframe en prenant de nouvelles variables (appelées axes, créés par l'ACP), qui condensent l'information en moins de variables, allégeant ainsi nos modèles de Machine Learning par la suite : on appelle cela de la réduction de dimension.

Pour les plus curieux, vous verrez également par la suite que l'ACP peut être très intéressante avant de réaliser un clustering…

Ressource : cours fourni « c-acp-afc-IUTSTID-sept2009.pdf »

Compétences

- Analyser une ACP

Consignes

- Utiliser le jupyter notebook 'Stat_descriptives_part3_ACP.ipynb', et suivre la trame proposée
- Répondre aux différentes questions

Focus sur les séries temporelles (1 jour)

Objectifs de l'activité

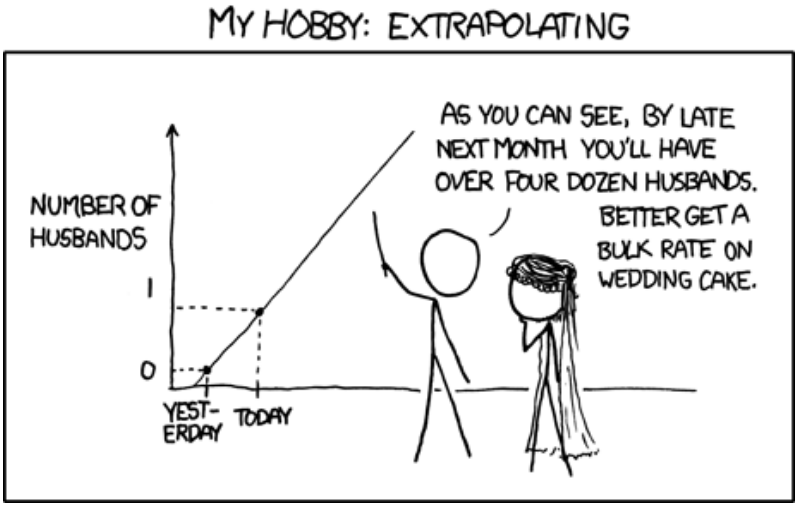
Une série temporelle est une série dont les observations sont mesurées à différents points dans le temps. Comme il s'agit d'évaluer l'évolution d'un phénomène au cours du temps, l'ordre des observations est important.

Les séries temporelles sont présentes dans de nombreux domaines d'application. Leur étude permet de comprendre les tendances passées ou prévoir les comportements futurs.

L'objectif de cette partie est donc de proposer des outils d'analyse exploratoire permettant de tirer une information utile de ces séries temporelles. Le dataset utilisé sera le même que pour la partie précédente.

La librairie Pandas contient des fonctions dédiées aux séries temporelles, qui simplifient énormément leur étude.

1. Utilisation des outils Pandas spécifiques aux séries temporelles
2. Identification de la structure de la série
3. Identifications de profils types
4. Dépendance temporelle (autocorrélation)
5. Études fréquentielles
6. Questions bonus !



Compétences

- Analyse d'une série temporelle
- En bonus (sans validation) : utilisation de Pandas pour l'étude des séries temporelles

Consignes

- Utiliser le jupyter notebook 'Stat_descriptives_part3_ST.ipynb', et suivre la trame proposée
- Répondre aux différentes questions