

# PROJET DATA – DU PYTHON

---

## CONTEXTE :

*Vous êtes un groupe de data scientist/analyst juniors travaillant pour une insurtech. Vous travaillez sur un projet d'exploitation de nouvelles bases de données d'assurance. Le seul outil à votre disposition est Python. Vous disposez de quelques semaines pour proposer une analyse aussi pertinente que possible des données. Vous rendez ensuite un rapport présentant toute l'étude à votre équipe et hiérarchie.*

## Consignes de fond

- Utilisez les différents outils, vus ensemble ou non, afin d'exploiter les données
- Identifiez les anomalies et retraitez les données du mieux possible : toutes les anomalies et les retraitements doivent être documentés
- Mettez en place des analyses et modélisations pertinentes : qui ont du sens et de l'intérêt, pas juste pour faire du python ...
- Interprétez vos résultats, déduire des relations entre variables et/ou observations, des tendances, etc.
- Proposez des analyses graphiques poussées (data visualisation) afin d'interpréter et comprendre les données
- Pour chacune des analyses et modélisations : il faut décrire l'objectif, ce que vous mettez en place pour y répondre ainsi que vos résultats : faire une étude sans l'explicitier ne sert strictement à rien !

*Petit rappel du processus classique d'analyse statistique :*

1. *Traitement de données : compréhension des données, analyse de la complétude et de la conformité des données, contrôles de cohérence (univarié/multivarié), gestion des anomalies : mise en place de correctif, exclusions, etc.*
2. *Jointures éventuelles de bases et contrôles*
3. *Analyses descriptives des données : univariée, multivariée, ASD*
4. *Analyse graphique (data visualisation) + Interfaçage via Shiny for Python*
5. *Modélisation : supervisée (régression, classification) vs non supervisée ; paramétrique (économétriques) vs non paramétriques (machine learning)*
6. *Analyse des résultats : interprétation, explications*
7. *Application : prévision, tarification, etc.*

## Consignes de forme

- Rédiger une note **propre** : via un support de votre choix mais à rendre en PDF, sans code.
- Rendre un script python **propre** : clair, structuré, précis et commenté !
- *Cent fotes dortograf est deux gramère bien évidemment*

## Consignes sur les modalités de rendu du projet

- Rappel : Votre projet « data » servira d'évaluation pour vos cours de « traitement de données », « Machine Learning » et « Econométrie ».

# PROJET DATA – DU PYTHON

---

- Vous avez jusqu'au **30/04 à 23h59 pour** nous déposer **vos 2 dossiers compressés** dans les répertoires suivants :
  - ✓ Partie Data+Econométrie : <https://amubox.univ-amu.fr/s/XbNM8stXELtkF5B>
  - ✓ Partie Machine Learning : <https://amubox.univ-amu.fr/s/bMA6jQkJ7csnKpe>
- fermeture des répertoires de dépôt le 06/05 – aucun projet ne sera pris en compte passé ce délai.
- 1 dossier par groupe : indiquez bien vos noms (nom du fichier ou dans le fichier) : groupe de 3p maximum, contenant :
  - ✓ *Une note explicative de votre travail : vos noms, une synthèse de vos résultats, objectifs, retraitements, votre choix de modélisation, etc.*
  - ✓ *Le code commenté et structuré (titres, sous-titres, etc.) avec vos résultats*
- Livrable : le script python et la note PDF
- Pas de dossier similaire : vous êtes autonomes et libre de faire ce que vous voulez donc vos dossiers seront forcément différents
- La note repose sur le fond et la forme de votre projet
- Données issues du papier : <https://arxiv.org/abs/2102.00252>
- ***Pas d'envoi par mail ! Un accusé de réception vous sera envoyé. Pénalité de retard.***