CONSIGNES POUR COMP_INVEST

Pour ce TP vous êtes invités à manipuler et exploiter un dataset en utilisant principalement les fonctionnalités apportées par panda.

Dans le fichier Excel Comp_Invest _Dataset.xlsx, vous trouverez 4 worksheets :

- COMPANY : Carte d'identité des entreprises
- INVESTMENT : Investissements reçus par les entreprises
- ACQUISITION : Acquisitions effectuées par les entreprises
- EMPLOYEE : Informations sur les employés des entreprises

Pour vous aider à mieux comprendre le dataset, on vous donne le « dictionnaire » ci-dessous :

Worksheet: COMPANY

Variable	Comment
COMPANY_NAME	Company name (unique ID)
CATEGORY	Industry category
LOCATION	Company location
FOUNDED_ON	Date that the company was founded
EXITED_ON	Date that the company exited (if any)
CLOSED_ON	Date that the company was closed (if any)
REVENUE_RANGE	Revenue range
EMPLOYEE_NUMBER	The number of employees

Worksheet: INVESTMENT

Variable	Comment
COMPANY_NAME	Company name
FUNDING_TYPE	The type of funding
MONEY_RAISED	The amount of money raised in the investment
ANNOUNCED_DATE	Date that the investment was announced
INVESTMENT_STAGE	The investment stage

Worksheet: ACQUISITION

Variable	Comment
COMPANY_NAME	Company name
ACQUIREE_NAME	Name of the acquired company
ANNOUNCED_DATE	Date that the acquisition was announced
PRICE	The price of acquisition
ACQUISITION_TYPE	The type of acquisition

Worksheet: EMPLOYEE

Variable	Comment
EMPLOYEE_MD5	Hashed unique ID for employee
JOB_TITLES	Job titles
COMPANY_NAME	Company name
ATTENDED_SCHOOLS	Schools that the employee has attended

Il n'y pas d'instructions fermes sur la nature du rendu. Vous pouvez utiliser toutes vos connaissances et techniques vues (en particulier relatives à panda) pour analyser le dataset et essayer d'en extraire des informations et à partir de vos observations d'en déduire des hypothèses.

Voici des questions types que vous pourriez vous poser :

- Combien d'entreprises ont reçu au moins une fois des investissements ?
- Quel type de distribution a-t-on de ces entreprises ? (Par géographie, Par taille, Par type d'industrie …)
- En moyenne, quelle est la somme qui a été levée par les entreprises au tour d'investissements de type Serie A ?
- Y a-t-il une corrélation entre le nombre de tours d'investissement ou le volume total d'investissement et la taille de l'entreprise ?
- Quelle entreprise a réalisé le plus grand nombre d'acquisitions ?
- Parmi ces entreprises, combien ont un CEO qui vient d'une « top school » ?

Les techniques que vous pourriez utiliser (mais ce n'est pas limitatif). :

- List / tuple / dictionary
- Pandas Series / DataFrame
- Data inspection
- Data cleaning
- Selecting and filtering
- str methods
- Summarization
- Sorting data
- Group by and apply
- Data merging
- Data reshaping
- Plotting

Votre rendu se fera sous la forme d'un notebook jupyter qui incluera vos résultats ainsi que des graphiques.

La note prendra en compte le fait que vous expliquerez les résultats obtenus, que vous commenterez les opérations effectuées (pour le cleaning et la transformation des dataframe en particulier), que vous utilisez bien quand c'est possible les méthodes apportées par panda, par la pertinence des questions que vous vous poserez, par l'esthétique et la pertinence des graphiques présentés.