Analyse descriptive avec Pandas

yzpt.github.io/module_pandas_b2

Attendus du module :

```
# Maîtriser la manipulation de données avec Pandas
- lecture et écriture de fichiers CSV
- sélection, filtrage, tri, groupement

# Nettoyer un jeu de données
- valeurs manquantes, suppression des doublons,

# Calculer et interpréter des statistiques descriptives avec Pandas
- tendance centrale, dispersion, distribution

# Explorer et analyser les relations entre plusieurs variables
- créer une matrice de corrélation
- analyse bi-variée scatter plots
- réaliser une régression linéaire simple

# Comprendre comment identifier et traiter les valeurs aberrantes avec Python
- méthode des z-scores
- IQR
- Box-plot
```

Mercredi 11 d&cemebre 2024

1. Cheatsheets:

- Pandas official cheatsheet
- EDA Cheatsheet
- Datacamp cheatsheet

2. Notebooks très bien rédigés :

Disponibles dans les dossiers 'workshop_notebooks' et 'cookbook': * Stefanie Molin Panda's wokrshop

3. Collection de ressources et datasets:

- ressources.md
- data/

Premiers notebooks & divers:

- 1.0.introduction.ipynb
- 2.0.gapminder_analysis.html

- 3.0.plotly_express.ipybn
- 4.0.outliers.ipynb
- 5.0.structure_pop.ipynb
- 6.0.join.ipynb

Jeudi 12 décembre 2024

1. Analyse de données Gapminder

- 2.0.gapminder_analysis.html
- 2.0.gapminder_analysis.ipynb

2. Template d'EDA à conserver/compléter

- 7.0.EDA_template.html
- 7.0.EDA_template.ipynb

3. Conversion de fichier HTML en notebook Très utile avec la fonctionanlité github Pages pour afficher des notebooks en ligne. - Capsule vidéo Github Pages

- 8.0.html_conversion.html
- 8.0.html_conversion.ipynb

4. Fichier clients_v2 à analyser : data/clients_v2.csv

Vendredi 13 décembre

1. Analyser les données des parkings de la MEL

- Etat des parkigns de la MEL:
- Informations : https://data.lillemetropole.fr/catalogue/dataset/disponibilite-parkings
- Depuis le fichier excel : data/parking.xls

Notebook : 12.0.parkings.ipynb HTML : 12.0.parkings.html

Consigne : Effectuer une analyse simple des données récupérées. * Donner un histogramme de l'occupation des parkings de la MEL. * Effectuer un classement des 5 parkings les plus occupés. * Effectuer une petite visualisation en groupant les parkings par ville.

Récupérer les données en temps réel au format JSON depuis une API :https://data.lillemetropole.fr/geoserver/wfs?SERVICE=WFS&REQUEST=GetFeature&VERSION=2.0.0&TYPENAMES=mel_mobilite_et_transport%3Aparking&OUTPUTFORMAT=application%2Fjson

2. **Outliers**

   - Faire une EDA du dataset 10.0.dataset_health.csv en détectant et éliminant les outliers.
   - Pour ce, s'appuyer sur le notebook 10.1.health_outliers.ipynb (html: 10.1.health_outliers.html)
   - Autre notebook d'exemples de traitement des outliers: divers/wine_outlie rs.ipynb

3. **S'exercer sur des datasets selectionnés :**

   - amazon_sales
   - animals
   - reddit_data
   - weather
   - data/iris.csv
   - data/titanic.csv
   - data/winequality.csv

Dataset du devoir surveillé :

- 11.0.dataset_orders.csv : dataset de commandes de produits
- 11.0.dataset_ogen_orders.ipynb : notebook d'exemple d'analyse (+ script de génération du dataset)

Effectuer une EDA qui couvre l'ensemble des points et l'envoyer par mail à l'adresse **yohann.zapart@gmail.com**, en-tête de sujet : [GEMA_B2]

Travail à rendre pour vendredi 20 décembre

- Effectuer une analyse descriptive sur un dataset de données réelles que vous choisierez.
- Le datset devra contenir au moins 1000 lignes et 10 colonnes (si ce n'est pas le cas, me demander par mail une homolgation).
- L'analyse devra contenir l'ensemble des points explicités dasn le fichier fiche.md