

Analyse de données avec python

Projet : Taux de chômage



Projet guidé Analyse de data de Thanksgiving

Ce mini projet portera sur les données du le taux de chômage pour des zones d'emplois correspondant à des villes de France. A chaque ville on associe une région. Ces données sont issue du site de l'INSEE.

Objectif : explorer les données et trouver des tendances ou hypothèses intéressantes.

Les données sont contenus dans le fichier « chomage_zone.xls ».

- Le fichier comporte 305 lignes et 73 colonnes.
- Les colonnes correspondes à :
 - ZE2010 : code zone d'emploie
 - LIBZE2010 : libellé zone d'emploie (ville)
 - REG : code région
 - LIBREG : libellé région
- Les 69 colonnes restantes correspondent au taux de chômage par années et par trimestre.

Introduction au dataset

- Lire le fichier « chomage_zone.xlsx » avec la librairie pandas et l'assigner à une variable data.
 - Ignorer les 5 première lignes.
 - Utiliser le noms des colonnes contenu dans la 1^{ère} ligne du fichier.
 - Le fichier contient plusieurs feuilles, sélectionner la bonne feuille à charger.
- Afficher les premières lignes du dataframe.
- Afficher le noms des colonnes avec l'attribut columns.

Filtrer les données

- Utiliser la méthode `Series.values_count()` pour afficher le décompte du nombre de zones d'emploies par département.
- Filtrer et garder toute les ligne du dataframe pour lesquelles le département est « Provence-Alpes-Côte d'Azur ».
- Assigner ce nouveau dataframe à `data2` et affiché le.
- Utiliser la méthode `Series.values_count()` pour afficher combien de fois une zone d'emploi apparait en « Provence-Alpes-Côte d'Azur ».

Exploration des taux de chômages

On cherche ici à savoir si le taux de chômage demeure élevé d'une année sur l'autre pour une même période.

- Créer un objet indiquant avec des booléens si le taux de chômage pour le 1ere trimestre 2020 est supérieur à 7.1
Assigner le résultat à la variable « data_t1_2020 ».
➤ Pour ce faire vous devrais utiliser « pd.eval() »
- Créer un objet indiquant avec des booléens si le taux de chômage pour le 1ere trimestre 2019 est supérieur à 7.1
Assigner le résultat à la variable « data_t1_2019 ».
- Créer un objet indiquant avec des booléens si le taux de chômage pour le 1ere trimestre 2018 est supérieur à 7.1
Assigner le résultat à la variable « data_t1_2018 ».
- Combiner les trois objets Series avec l'opérateur « & » et assigné le résultat à la variable « data_7_1 ».
- data_7_1 est une liste. Assigné là à un dataframe dont le nom de colonne sera « booléen ».
- Afficher les valeurs unique et combien de fois elle apparaissent dans la colonnes.

Lien entre taux de chômage et nombre de pôle emploi

- Charger le fichier « pole_emploi.xlsx » après l'avoir ouvert et visualiser et l'assigner à la variable « p_emploi ».
 - Vous ne pouvez pas utiliser les noms de colonnes contenus dans le fichier.
- Joindre les data et p_emploi au sein d'un seul et unique dataframe en faisant correspondre les régions par leur codes. Utiliser `pd.merge()` pour réaliser cette jointure.
- Créer une colonne quantile dans p_emploi qui prendra les valeurs ['q25', 'q50', 'q75', 'q100'] en fonction du quartile auquel appartient le nombre de pôle emploi. Pour se faire utiliser « `pd.qcut()` ».
- Générer un pivot de table montrant la moyenne du taux de chômage du trimestre 1 de l'année 2020 pour chaque régions et quantile du nombre de pôle emploi.
 - Appeler la méthode `pivot_table()` sur le data frame
 - Passer au paramètre « index » la valeur « LIBREG ».
 - Passer au paramètre « columns » la valeur « quantile ».
 - Passer au paramètre « values » la valeur « 2020-T1 »
 - Afficher les résultats.