Master Spécialisé « Data Engineering » (MSDE)

Projet Module 3 :

Programmation Python pour Data Science

Objectifs et données :

Le projet porte sur le traitement d'un jeu de données du recensement général de la population et de l'habitat de 2004 (RGPH2004) de la région Marrakech-Tensift-Al Haouz selon le découpage administratif de cet époque (un nouveau découpage administratif a été adopté en 2009).

L'objectif de ce projet est de traiter et d'analyser les données démographiques des communes de cette région à travers le langage python et ses librairies afin de répondre aux questions posées ci-dessous.

Les données sont présentées dans le fichier Excel (première feuille du fichier Recensement_RGPH2004_RegionMTA.xlsx) joint à l'énoncé du projet et qui donne une série de caractéristiques démographiques pour les communes de cette région.

La description des champs de ce tableau est présentée dans la deuxième feuille de ce fichier Excel.

Un autre fichier auxiliaire donnant le code et le nom des provinces de la région est fourni sous format CSV (Code_Nom_Province.csv).

Deadline de remise du projet : 12 janvier 2019

Livrables du projet :

Fichier zippé (.zip ou .rar) portant votre (vos) noms et comportant les fichiers suivants :

- Fichier notebook complet contenant le code et les résultats du code (.ipynb) y compris les graphiques
- Fichiers de sorties mentionnés dans les questions 15, 21 et 29 (*à mettre dans le même répertoire que le fichier ipynb*)

Remarques :

- Le projet peut être réalisé en monôme ou en binôme

- Les livrables sont à déposer sur la plateforme du Master dans l'espace du module 3 : **Dépôt du projet python**
- Les graphiques à produire doivent être inclus dans le fichier notebook (.ipynb) «mettre le paramètre suivant : %matplotlib inline». Pour chaque série de données, vous devez choisir le type de graphique le plus adéquat et en spécifiant le titre du graphique, les titres des axes et sa légende.

Questions :

Partie 1 (24 points) :

1. Créer votre fichier notebook et charger dans un DataFrame (DataFrame des commune : dfc) le contenu du fichier Excel (Recensement_RGPH2004_RegionMTA.xlsx)
2. Afficher les dix premières lignes et dix dernières lignes de « dfc »
3. Calculer les statistiques (min, max, moyenne, médian, somme et écart-type) des champs « population » et « Surface_ha »
4. Supprimer la colonne « ID » du DataFrame « dfc »
5. Combien y-a-t-il de valeurs uniques dans le champ « Nom_Commune » et déduire les valeurs qui se répètent plus d'une fois et leur fréquence.
6. Quelle est la commune ayant le taux d'activité le plus bas et celle ayant le taux d'activité le plus haut
7. Quelles sont les communes ayant simultanément une population supérieure à 10000 et un taux d'analphabétisme supérieur à 40%
8. Quelles sont les communes dont le nom commence ou se termine par la lettre « A ».
9. Représenter graphiquement la répartition de la population des communes (*voir le volet remarques ci-dessus).* Commenter le graphique.
10. Représenter graphiquement le taux d'activité en fonction du taux d'analphabétisme. Commenter le graphique.
11. Ajouter dans le « dfc » une nouvelle colonne appelée « densite_pop » représentant la densité de la population des communes (unité : hab/km2 cad nombre d'habitants / kilomètre carré)
12. Ecrire une fonction Python appelée « codeProvince » qui permet d'extraire le code de la province à partir du code de la commune. [*le code de la province est représenté par les trois premiers caractères du code de la province*]

Exemple :

Code_Commune	Code_Province
041.03.01	041
211.05.01	211

Ajouter la colonne « Code_Province » aux données « dfc » en appliquant la fonction « codePronvice » sur « dfc » *(pour les codes communes commençant par zéro, il faut générer un code province commençant par zéro. par exemple « 041 » au lieu de « 41»)*

13. Ecrire une fonction Python appelée « nomProvince » qui permet de déterminer le nom de la province à partir de son code en lisant un fichier CSV comportant ces deux informations (*fichier Code_Nom_Province.csv).*
    Appliquer cette fonction sur « dfc » pour ajouter la colonne « Nom_Province » aux données « dfc »
14. Mettez comme index pour « dfc » les deux colonnes « Code_Province » et « Code_Commune » puis trier le selon ce nouvel index.
15. Exporter votre DataFrame « dfc » vers un fichier Excel portant le nom « dfc_votreNom.xlsx »
16. Calculer le nombre de communes rurales et urbaines de chaque province
17. Calculer le taux d'activité (*population active/population totale*) de chaque province
18. Générer un nouveau DataFrame (DataFrame des Provinces : dfp) qui donne les informations suivantes pour chaque province :
    - Nom de la province
    - Code de la province
    - Surface de la province
    - Population de la province
    - Population active de la province
19. Calculer le taux d'activité et le taux d'analphabétisme par province et ajouter les dans « dfp »
20. Calculer les pourcentages de répartition de population par tranche d'âge (Pop_inf6_pc, Pop_615_pc, Pop_1560_pc, Pop_sup60_pc) pour chaque province et ajouter les dans « dfp »
21. Exporter votre DataFrame « dfp » vers un fichier Excel portant le nom « dfp_votreNom.xlsx »
22. Représenter sur le même graphique la population et la population active des provinces. Commenter le graphique.
23. Représenter sur le même graphique le taux d'activité et le taux d'analphabétisme des provinces. Commenter le graphique.
24. Représenter sur le même graphique les pourcentages de répartition de population par tranche d'âge (Pop_inf6_pc, Pop_615_pc, Pop_1560_pc, Pop_sup60_pc). Commenter le graphique.

Partie 2 (6 points) :

25. Mettez en place une démarche automatique en utilisant python et ses librairies pour extraire les informations citées ci-dessous (caractéristiques démographiques des communes en 2014) pour chacune des communes représentées dans le DataFrame « dfc » à partir du site web du Haut-Commissariat au plan (http://rgphentableaux.hcp.ma/). Appeler ce nouveau DataFrame « dfc2 »
    - Population de la commune
    - Age moyen au premier mariage
    - Taux d'activité
    - Taux d'analphabétisme

    « dfc2 » contient les informations du « dfc » plus les nouvelles informations de 2014.

26. Calculer le taux d'accroissement de la population de chaque commune ((population 2014-Population *2004)/population 2004*)
27. Calculer le taux d'accroissement de la population de chaque province (*on garde toujours le même découpage administratif tel que défini dans le fichier Excel initial : Recensement_RGPH2004_RegionMTA.xlsx)*
28. Tracer un graphique qui met en exergue la variation de la population des provinces entre 2004 et 2014
29. Exporter votre DataFrame « dfc2 » vers un fichier Excel portant le nom « dfc2_votreNom.xlsx »