Analyse de données avec python

Series et Dataframe

Structure clé de Pandas

3 types d’objets dans Pandas :

- Series
- Dataframes : ensemble d’objets Series
- Panels : Ensemble d’objets Dataframe

Les Series

- Correspond à une colonne. Plusieurs Series contenu dans un objet forme un Dataframe.
- Utilises les valeurs NaN pour gérer les valeurs manquantes
- Types de données :
  - Float
  - Int
  - Bool
  - Datetime64[ns] : Date et horaire sans la time zone
  - Datetime[ns, tz] : Date et horaire avec la time zone
  - Timedelta[ns] : Différence de date et horaire (seconde, minute, …)
  - Category : pour les variables catégorielles
  - Object : chaîne de caractère.

Projet guidé Analyse de data de Thanksgiving

Ce mini projet portera sur les résultats d'un sondage ayant pour sujet ce que mange les Américains au moment de Thanhsgiving, leurs revenus moyen entre autre choses.

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

Les données sont contenus dans le fichier « thanhsgiving.csv ».

- Le fichier comporte 1059 lignes et 65 colonnes.
- La 1ère ligne correspond aux questions posées et pourrons servir de noms des colonnes.
- La première colonne contient un id pour chaque personne interrogé.
- Pour de nombreuse questions les réponses sont catégorielles (plusieurs choix de réponses possible).

Introduction au dataset

- Lire le fichier « thanksgiving.csv » avec la librairie pandas et l'assigner à une variable data.
  - Spécifier dans les paramètre de la fonction permettant de lire le fichier « encoding='latin-1' » car ce dataset n'est pas encodé normalement.
  - Utiliser le noms des colonnes contenu dans la 1ère ligne du fichier.
- Afficher les premières lignes du dataframe (une méthode en particulier permet de le faire).
- 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 réponses pour chacune des modalités de la colonnes « Do you celebrate Thanksgiving? »
- Filtrer et garder toute les ligne du dataframe pour lesquelles la réponse à la question « Do you celebrate Thanksgiving? » est « Yes ».
- Assigner ce nouveau dataframe à data et affiché le.

Exploration des repas de Thanksgiving

- Utiliser la méthode Series.values_count() pour afficher combien de fois chaque résultats apparait pour la question « What is typically the main dish at your Thanksgiving dinner? »
- Afficher la colonne « Do you typically have gravy? » pour les ligne du dataframe data pour lesquelles la colonne « What is typically the main dish at your Thanksgiving dinner? » vaut « Tofurkey » pour la dinde de tofu.

Exploration des desserts pour Thanksgiving

On cherche ici à savoir combien de personnes ont consommés des tartes à la pomme, la citrouille ou pécan.

- Créer un objet Series indiquant avec des booléens les valeurs de la colonnes « Which type of pie is typically served at your Thanksgiving dinner? Please select all that apply. - Apple » qui sont nulles.
  Assigner le résultat à la variable « apple_isnull ».
- Créer un objet Series indiquant avec des booléens les valeurs de la colonnes « Which type of pie is typically served at your Thanksgiving dinner? Please select all that apply. - Pumpkin » qui sont nulles.
  Assigner le résultat à la variable « pumpkin_isnull ».
- Créer un objet Series indiquant avec des booléens les valeurs de la colonnes « Which type of pie is typically served at your Thanksgiving dinner? Please select all that apply. - Pecan » qui sont nulles.
  Assigner le résultat à la variable « pecan_isnull ».
- Combiner les trois objets Series avec l'operateur « & » et assigné le résultat à la variable « pies ».
- Afficher les valeurs unique et combien de fois elle apparaissent dans la colonnes de pies.

Convertir l’âge en valeur numérique

- Ecrire une fonction qui converti une chaîne de caractère en une valeur entière. Cela permettra de convertir les valeurs de la colonne « Age » en entiers. Cette fonction prendra en paramètre une chaîne de caractères (les valeurs actuelles de la colonne « Age »)
  - Utiliser la fonction is_null() pour vérifier si les valeurs sont nulles. Ajouter une condition if qui retourne None si la valeur est nulle.
  - Séparer les chaine de caractère en fonction de l’espace (‘ ’) et extraire le 1ère élément de la liste.
  - Supprimer le caractère ‘+’ dans le résultat.
  - Convertir le résultat en entier.
  - Retourner le résultat.
- Utiliser la méthode Series.apply() pour appliquer la fonction à chaque valeur de la colonne ‘Age’ du dataframe data.
  - Assigner le résultat à la nouvelle colonne ‘int_age’ du dataframe.
- Appeler la méthode Series.describe() sur la colonne « int_age » du dataframe data et afficher le résultat.

Convertir les revenus en valeurs numérique

- Ecrire une fonction pour convertir les revenus en valeur unique de format entier.
  - Utiliser la fonction isnull() pour vérifier si la valeur est nulle. Si c'est le cas, retourner « None ».
  - Séparer la chaine de caractère en prenant l'espace comme délimiteur et extraire le premier élément de la liste résultante.
  - Si le résultat vaut « Prefer » retourner « None ».
  - Supprimer les caractères « $ » et « , ».
  - Utiliser int() pour convertir le résultat en entier.
  - Retourner le résultat.
- Utiliser la méthode Series.apply() pour appliquer la fonction précédente à chaque valeur de la colonne « How much total combined money did all members of your HOUSEHOLD earn last year? » du dataframe data.
  - Assigner le résultat à la nouvelle colonne « int_income » du dataframe data.
- Appeler la méthode Series.describe() à la colonne int_income du dataframe data et afficher le résultat.

Lien entre distance et revenus

- Regarder de quel manière les personnages gagnant moins de 150 000 dollars voyagent.
  - Filtrer data en sélectionnant seulement les valeur de « int_income » infèrieures à 150 000
  - Sélectionner la colonne « How far will you travel for Thanksgiving? » en prenant en compte le filtre.
  - Utiliser la méthode value_counts() pour compter combien e fois chaque vaaleur apparait dans la colonne.
  - Afficher le résultats.
- Faire de même avec les personnages gagnant plus de 150 000 dollars.

Lien entre passer Thankgiving entre amis avec l’âge et le revenus

- Générer un pivot de table montrant la moyenne d’âge des sondés pour chaque catégorie des questions « Have you ever tried to meet up with hometown friends on Thanksgiving night? » et « Have you ever attended a "Friendsgiving?" ».
  - Appeler la méthode pivot_table() sur le data frame data.
  - Passer au paramètre « index » la valeur « Have you ever tried to meet up with hometown friends on Thanksgiving night? ».
  - Passer au paramètre « columns » la valeur « Have you ever attended a "Friendsgiving?" ».
  - Passer au paramètre « values » la valeur « int_age »
  - Afficher les résultats.
- Faire de même avec les revenus avec ces deux questions.