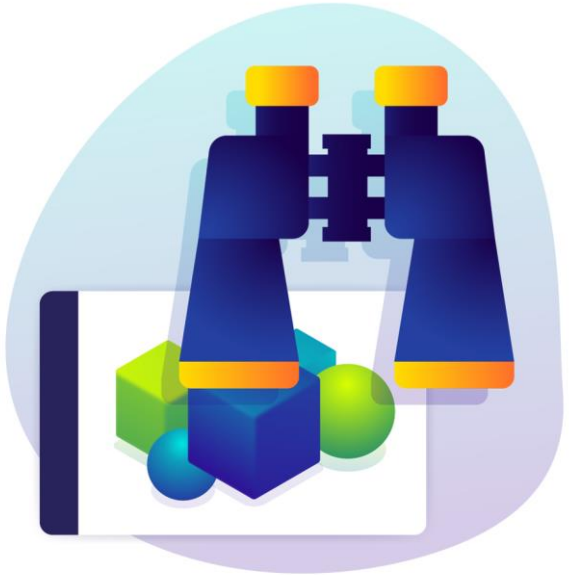
Projet guidé: Analyse de data

Sondage Thanksgiving

Partie 1 : filtrer et explorer les tendance



Introduction au dataset

- Utiliser la fonction pandas.read_csv() pour lire le fichier thanksgiving.csv.
  - Spécifier dans les paramètres de la fonction l'argument encoding='latin-1' car ce dataset n'est pas encodé normalement
  - Assigner le dataframe à la variable data
- Afficher les premières lignes de data.
- Afficher le nom des colonnes avec l'attribut columns.

Filtrer des données

- Utiliser la méthode Series.value_counts() pour afficher le décompte du nombre de réponses pour chaque catégorie de la colonne 'Do you celebrate Thanksgiving?'.
- Filtrer et garder toutes les lignes du dataframe data pour lesquelles la réponse à la question 'Do you celebrate Thanksgiving?' est 'Yes'.
- Assigner ce nouveau dataframe à data et afficher le.

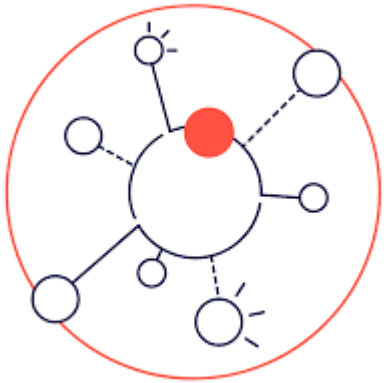
Exploration des repas de thanksgiving

- Utiliser la méthode Series.value_counts() pour afficher combien de fois chaque résultat apparaît pour la question et colonne 'What is typically the main dish at your Thanksgiving dinner?'.
- Afficher la colonne 'Do you typically have gravy?' pour les lignes du dataframe data pour lesquelles la colonne 'What is typically the main dish at your Thanksgiving dinner?' vaut 'Tofurkey' pour la dinde de tofu.
  - Créer un filtre qui sélectionne seulement les lignes du dataframe data où 'What is typically the main dish at your Thanksgiving dinner?' vaut 'Tofurkey'
  - Assigner le résultat à la variable tofurkey
  - Afficher la colonne 'Do you typically have gravy?' de ce nouveau dataframe tofurkey

Tendances des desserts pour Thanksgiving

- Créer un objet Series indiquant avec des booléens les valeurs de la colonne '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 colonne '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 colonne '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 ces 3 objets Series avec l'opérateur & et assigner le résultat à la variable pies.
- Afficher les valeurs uniques et combien de fois elles apparaissent dans la colonne de pies.

Partie 2 : Liens et corrélations entre variables



Convertir l'âge en une valeur numérique

- Ecrire une fonction qui convertit une chaine de caractères en une valeur entière. Cela permettra de convertir les valeurs de la colonne 'Age' en entiers. Cette fonction prendra en paramètre une chaine de caractères (les valeurs actuelles de la colonne 'Age')
  - Utiliser la fonction isnull() pour vérifier si les valeurs sont nulles. Ajouter une condition if qui retourne 'None' si la valeur est nulle
  - Séparer les chaines de caractères en fonction de l'espace (' ') et extraire le premier élément de la liste résultante (méthode split())
  - Remplacer le caractère '+' dans le résultat avec une chaine de caractères vide en remplacement pour le supprimer (méthode replace())
  - utiliser int() pour 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 data
- Appeler la méthode Series.describe() sur la colonne 'int_age' du dataframe data et afficher le résultat.

Convertir les revenus en valeurs numériques

- 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 caractères en prenant l'espace comme délimiteur et extraire le premier élément de la liste résultante
  - Si le résultat vaut 'Prefer', retourner 'none'
  - Remplacer les caractères '$' et ',' par des chaines de caractères vides pour les supprimer
  - 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.

Corrélation entre distance et revenus

- Regarder de quelle manière les personnes gagnant moins de 150000 dollars voyagent.
  - Filtrer data en sélectionnant seulement les valeurs de 'int_income' inférieures à 150000
  - Sélectionner la colonne 'How far will you travel for Thanksgiving?' en prenant en compte le filtre
  - Utiliser la méthode value_counts() pour compter combien de fois chaque valeur apparaît dans la colonne
  - Afficher les résultats
- Faire de même avec les personnes gagnant plus de 150000 dollars.

Lien entre passer Thanksgiving entre amis avec l'âge et le revenu

- 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 dataframe 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 pour les revenues avec ces 2 questions.