Zaawansowane ćwiczenia z biblioteki `pandas`

Wprowadzenie

Ćwiczenia te maja na celu rozwiniecie umiejetności w korzystaniu z biblioteki `pandas` do zaawansowanej analizy danych. Pracujemy na dwóch plikach CSV: `astronomical_data_file1.csv` i `astronomical_data_file2.csv`, które zawieraja dane obserwacyjne.

Ćwiczenia

1. **Wczytaj dane i zweryfikuj strukture.** Wczytaj oba pliki CSV do obiektów `DataFrame` i zweryfikuj ich struktury, wyświetlajac podstawowe informacje o kolumnach oraz brakujacych wartościach.
2. **Porównaj rozkład brakujacych danych.** Przedstaw graficznie rozkład brakujacych danych w obu plikach, wykorzystujac np. wykres słupkowy.
3. **Połacz dane w jeden DataFrame.** Połacz oba zestawy danych, zachowujac informacje o pochodzeniu każdej kolumny poprzez dodanie odpowiednich sufiksów.
4. **Zidentyfikuj wspólne i unikalne obiekty.** Znajdź obiekty (na podstawie `ObjectID`), które wystepuja tylko w jednym z plików oraz te, które sa wspólne dla obu plików.
5. **Oblicz mediane i odchylenie standardowe.** Oblicz mediane i odchylenie standardowe dla kolumn `Brightness` i `Distance` w obu plikach, zarówno dla obiektów wspólnych, jak i unikalnych.
6. **Uzupełnij brakujace wartości metoda interpolacji.** Uzupełnij brakujace wartości metoda interpolacji liniowej, a nastepnie porównaj wyniki ze średnimi oryginalnych kolumn.
7. **Znajdź obiekty o najwyższych i najniższych wartościach.** W obu plikach znajdź obiekty o skrajnych wartościach jasności (`Brightness`) i predkości (`Velocity`) i sprawdź, czy pojawiaja sie w obu plikach.
8. **Sprawdź korelacje miedzy kolumnami.** Przeprowadź analize korelacji miedzy kolumnami (np. `Brightness`, `Distance`, `Velocity`) i przedstaw wyniki w formie macierzy korelacji.

9. **Przeanalizuj zmienność jasności i odległości.** Podziel obiekty na grupy według przedziałów jasności i odległości, a nastepnie oblicz średnia i odchylenie standardowe dla każdej grupy.

10. **Przefiltruj dane na podstawie warunków.** Znajdź wszystkie obiekty, które maja jasność powyżej 15, odległość poniżej 5 oraz predkość powyżej 100. Przedstaw wyniki na wykresie.

11. **Zidentyfikuj trend w danych.** Utwórz wykres pokazujacy, jak jasność zmienia sie wraz z odległościa dla obiektów wspólnych dla obu plików. Dodaj linie trendu.

12. **Oblicz różnice dla wspólnych obiektów.** Dla obiektów obecnych w obu plikach, oblicz różnice w jasności i predkości miedzy plikami. Zidentyfikuj obiekty o najwiekszych różnicach.

13. **Grupowanie danych według kolumny.** Grupuj obiekty według ich jasności i oblicz statystyki dla każdego przedziału jasności, m.in. średnia odległość i średnia predkość.

14. **Porównaj zmienne miedzy plikami.** Zrób porównanie rozkładów jasności, odległości i predkości miedzy plikami za pomoca wykresu pudełkowego (boxplot).

15. **Analiza wielkości brakujacych danych.** Zidentyfikuj kolumny o najwiekszej liczbie brakujacych wartości w każdym pliku. Przedstaw wykres z rozkładem braków.

16. **Wyznacz wartości odstajace.** Znajdź wartości odstajace w kolumnach `Brightness` i `Distance` na podstawie zdefiniowanych progów. Wyświetl je i wskaż, czy wystepuja w obu plikach.

17. **Wypełnij wartości brakujace średnia lub mediana.** Uzupełnij brakujace wartości w kolumnach `Brightness` i `Distance` na podstawie ich średniej lub mediany. Porównaj wyniki obu metod.

18. **Eksport przetworzonych danych.** Po wykonaniu wszystkich powyższych analiz, zapisz przetworzone dane w trzech plikach: wszystkie dane, dane wspólne i dane unikalne.

19. **Tworzenie funkcji analizy danych.** Stwórz funkcje, która automatycznie analizuje dane wejściowe i wykonuje operacje takie jak wypełnianie braków, wyznaczanie statystyk i znajdowanie wspólnych obiektów.

20. **Porównanie za pomoca wykresów.** Użyj wykresów histogramów, aby porównać rozkład wartości `Brightness` oraz `Velocity` miedzy obiektami wspólnymi i unikalnymi w obu plikach.

Podpowiedzi

Do wykonania powyższych zadań pomocne moga być nastepujace funkcje biblioteki `pandas`:

- `pd.read_csv()` - wczytywanie danych z plików CSV,
- `df.info()` - uzyskanie informacji o strukturze danych,
- `df.isna()` i `df.dropna()` - sprawdzanie i usuwanie brakujacych wartości,
- `df.fillna()` - wypełnianie brakujacych wartości,
- `pd.merge()` - łaczenie zbiorów danych na podstawie klucza,
- `df.groupby()` - grupowanie danych,
- `df.corr()` - obliczanie korelacji miedzy kolumnami,
- `df.describe()` - wyznaczanie podstawowych statystyk,
- `df.median()` i `df.mean()` - obliczanie mediany i średniej,
- `df.plot()` i `df.hist()` - tworzenie wykresów,
- `df.duplicated()` - znajdowanie duplikatów,
- `df.apply()` - stosowanie funkcji do kolumn,
- `df.to_csv()` - eksport przetworzonych danych do plików CSV.