Pandas - zadania

Zadanie o samolotach (nycflights13)

Potrzebne pliki można pobrać np. z https://github.com/gagolews/Analiza_danych_w_jezyku_Python/tree/master/zbiory_danych/nycflights13

Baza danych zawiera:

1. `flights` - informacje o lotach,
2. `airports` - nazwy i położenia lotnisk,
3. `planes` - informacje o samolotach,
4. `airlines` - nazwy linii lotniczych,
5. `weather` - godzinowe dane meteorologiczne

Kod potrzebny do zadania

```
import numpy as np
import pandas as pd
import sqlite3

airlines = ... # tu wczytaj plik csv
airports = ... # tu wczytaj plik csv
flights = ... # tu wczytaj plik csv
planes = ... # tu wczytaj plik csv
weather = ... # tu wczytaj plik csv

con = sqlite3.connect('nycflights13.db')

planes.to_sql("planes", con, if_exists="replace")
airlines.to_sql("airlines", con, if_exists="replace")
flights.to_sql("flights", con, if_exists="replace")
airports.to_sql("airports", con, if_exists="replace")
weather.to_sql("weather", con, if_exists="replace")
```

Zapytanie, żeby przetestować, czy działa:

```
sql_result = pd.read_sql_query("SELECT * FROM planes", con)
sql_result
```

Operacje analogiczne do zapytań SQL

Zadanie polega na wykonaniu z użyciem biblioteki Pandas operacji na ramkach danych, które dadzą nam wyniki analogiczne do następujących komand *SQL*. Wynikiem powinna być zawsze ramka danych.

1. `SELECT DISTINCT engine FROM planes`
2. `SELECT DISTINCT type, manufacturer FROM planes`
3. `SELECT COUNT(*), engine FROM planes GROUP BY engine`
4. `SELECT COUNT(*), engine, type FROM planes GROUP BY engine, type`
5. `SELECT MIN(year), AVG(year), MAX(year), engine, manufacturer FROM planes GROUP BY engine, manufacturer`
6. `SELECT * FROM planes WHERE speed IS NOT NULL`

7. `SELECT tailnum FROM planes WHERE year >= 2010`
8. `SELECT tailnum FROM planes WHERE seats BETWEEN 100 and 200 LIMIT 20`
9. `SELECT * FROM planes WHERE manufacturer IN ("BOEING", "AIRBUS", "EMBRAER")`
10. `SELECT * FROM planes WHERE manufacturer IN ("BOEING", "AIRBUS", "EMBRAER") AND seats>300`
11. `SELECT manufacturer, COUNT(*) FROM planes WHERE seats > 200 GROUP BY manufacturer`
12. `SELECT manufacturer, COUNT(*) FROM planes GROUP BY manufacturer HAVING COUNT(*) > 10`
13. `SELECT manufacturer, COUNT(*) FROM planes WHERE seats > 200 GROUP BY manufacturer HAVING COUNT(*) > 10`
14. `SELECT manufacturer, COUNT(*) AS howmany FROM planes GROUP BY manufacturer ORDER BY howmany`
15. `SELECT manufacturer, COUNT(*) AS howmany FROM planes GROUP BY manufacturer ORDER BY howmany DESC LIMIT 10`
16. `SELECT * FROM planes WHERE year >= 2012 ORDER BY year, seats`
17. `SELECT * FROM planes WHERE year >= 2012 ORDER BY seats, year`

Inne

18. Wybierz 100 losowych wierszy z airports,
19. Wybierz 5% losowych wierszy,
20. Wybierz pierwszych 100 wierszy,
21. Wybierz ostatnich 100 wierszy.

Teoriomnogościowe

Niech:

- A - wiersze 1,. . . ,10 z airports,
- B - wiersze 6,. . . , 15 z airports.

1. `SELECT * FROM A UNION SELECT * FROM B`
2. `SELECT * FROM A UNION ALL SELECT * FROM B`
3. `SELECT * FROM A INTERSECT SELECT * FROM B`
4. `SELECT * FROM A EXCEPT SELECT * FROM B`
5. `SELECT * FROM B EXCEPT SELECT * FROM A`

Join

Na koniec dokonaj bazodanowej operacji join:

1. Złącz flights z planes
2. Złącz flights z airports
3. Złącz flights z weather
4. Złącz flights z weather, planes i airports

Zadanie o filmach (MovieLens)

Potrzebne pliki można pobrać np. z http://grouplens.org/datasets/movielens/1m/

Następnie należy zaznajomić się z danymi czytając plik `README`.

0. Jak jest przechowywany wiek? Jak są przechowywane zawody wykonywane przez użytkowników?

1. Wczytaj wszystkie trzy pliki jako osobne ramki danych do Pandasa.
2. Stwórz nową kolumnę `Year` w ramce danych, która zawierać będzie rok powstania filmu (jako `int`).
3. Ile jest wszystkich filmów?
4. Wykryj i wypisz wszystkie remaki (ten sam tytuł, inny rok produkcji)
5. Kontynuacja poprzedniego, o remakach: jaki film ma największą odległość od najstarszej wersji do najnowszej?
6. Ile filmów powstało w poszczególnych latach
7. Jak wygląda rozkład płci oraz grup wiekowych wśród użytkowników (zarówno jednowymiarowo, jak i dwuwymiarowo)
8. Jaki gatunek filmowy jest najczęstszy
9. Jaki jest najlepszy film wszechczasów, (najlepszy, czyli ma najwyższą średnią ocenę) - to zadanie możesz rozwiązać wykonując złączenie (join) zbioru movies i ratings
10. Wykonaj poprzedni punkt, odrzucając wcześniej filmy które nie uzyskały wystarczająco dużo głosów (np 100)
11. Jaki jest najlepszy film według kobiet i według mężczyzn
12. Jaki jest średni rok oglądanego filmu w poszczególnych grupach wiekowych
13. Jakie trzy gatunki filmowe są najczęściej oglądane przez kobiety i mężczyzn