

2. Pandas

2.1 Ramka danych flights zawiera kolumny year i month. Wygeneruj na ich podstawie nową zmienną typu data i czas.

2.2 Zastosuj wektor etykiet hierarchicznych do nazwania wierszy ramki danych flight. Wykorzystaj do tego celu dane zawarte w kolumnach year i month.

2.3 Podziel ramkę danych iris na dwie rozłączne części: niech pierwsza zawiera 80% obserwacji, a druga - pozostałe 20%. (podpowiedź: sample albo permutation)

2.4 Dokonaj standaryzacji wszystkich zmiennych liczbowych w ramce danych iris (odejmij średnią i podziel przez odchylenie standardowe).

2.5 Utwórz tabelę przestawną (pivot table) dla ramki tips, dla wartości total_bill i tip, w rozbiciu na dzień tygodnia oraz płeć. Funkcją agregującą niech będzie mediana.

2.6 Utwórz tablicę kontyngencji dla kombinacji zmiennych tips.smoker, tip.sex, tips.day.

2.7 Oblicz, jaką część rachunku stanowi napiwek i podaj podsumowania liczbowe tych wartości (także w podgrupach utworzonych na podstawie zmiennych jakościowych dostępnych w ramce danych tips)

2.8 Wyznacz podstawowe statystyki próbkowe (średnia, odchylenie standardowe, kwartyle) dla liczby pasażerów (ramka danych flights) w każdym roku z osobna. Przedstaw wyniki tak, by zagregowane wartości przechowywane były w kolumnach, a kolejne lata - w wierszach.

2.9 Niech `flights_pivot = flights.pivot("month", "year", "passengers")`
Odwróć powyższe przekształcenie, tzn. przekształć macierz `flights_pivot` na ramkę danych tożsamą z `flights`.

2.10

Pobierz ramki danych z

https://github.com/gagolews/Analiza_danych_w_jezyku_Python/tree/master/zbiory_danych/nycflights13 lub mojej strony

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

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, model, seats, engine 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`
-
1. `SELECT * FROM planes WHERE speed IS NOT NULL`
 2. `SELECT tailnum FROM planes WHERE year >= 2010`

3. SELECT tailnum FROM planes WHERE seats BETWEEN 100 and 200 LIMIT 20
4. SELECT * FROM planes WHERE manufacturer IN ("BOEING", "AIRBUS", "EMBRAER")
5. SELECT * FROM planes WHERE manufacturer IN ("BOEING", "AIRBUS", "EMBRAER") AND seats>300

1. SELECT manufacturer, COUNT(*) FROM planes WHERE seats > 200 GROUP BY manufacturer
2. SELECT manufacturer, COUNT(*) FROM planes GROUP BY manufacturer HAVING COUNT(*) > 10
3. SELECT manufacturer, COUNT(*) FROM planes WHERE seats > 200 GROUP BY manufacturer HAVING COUNT(*) > 10

1. SELECT manufacturer, COUNT(*) AS howmany FROM planes GROUP BY manufacturer ORDER BY howmany
2. SELECT manufacturer, COUNT(*) AS howmany FROM planes GROUP BY manufacturer ORDER BY howmany DESC LIMIT 10
3. SELECT * FROM planes WHERE year >= 2012 ORDER BY year, seats
4. SELECT * FROM planes WHERE year >= 2012 ORDER BY seats, year

1. Wybierz 100 losowych wierszy z airports,
2. Wybierz 5% losowych wierszy,
3. Wybierz pierwszych 100 wierszy,
4. Wybierz ostatnich 100 wierszy.

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

Tutaj należy dokonać 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

2.11 Należy pobrać zbiór danych Movie Lens spakowany w pliku ml-1m.zip dostępny tutaj: <http://grouplens.org/datasets/movielens/1m/> lub mojej strony.

Następnie należy zaznajomić się z danymi czytając plik README. Później należy wczytać wszystkie trzy pliki jako osobne ramki danych do Pythona. Zwróć uwagę na separator i zastanów się jak rozwiązać ten problem. Po pomyślnym wczytaniu danych stwórz nową kolumnę w ramce danych, która zawierając będzie rok powstania filmu. Następnie odpowiedz na poniższe pytania

- A. ile jest wszystkich filmów
- B. ile filmów powstało w poszczególnych latach
- C. jak wygląda rozkład płci oraz grup wiekowych wśród użytkowników
- D. jaki gatunek filmowy jest najczęstszy
- E. jaki jest najlepszy film wszechczasów, (najlepszy, czyli ma najwyższą średnią ocenę) - to zadanie możesz rozwiązać wykonując złączenie (join) zbioru movies i ratings
- F. wykonaj poprzedni punkt, odrzucając wcześniej filmy które nie uzyskały wystarczająco dużo głosów (np 100)
- G. jaki jest najlepszy film według kobiet i według mężczyzn

H. jaki jest średni rok oglądanego filmu w poszczególnych grupach wiekowych

I. jakie trzy gatunki filmowe są najczęściej oglądane przez kobiety i mężczyzn