PDU 2022/2023

Praca domowa nr 4 - Projekt (max. = 30 p.)

Zadanie rozwiązujemy w grupach dwu- lub trzyosobowych.

Prace domowe należy przesłać za pośrednictwem platformy `Moodle` (jedna osoba z zespołu) – **jedno archiwum `.zip`**[1] o nazwie typu `Nazwisko1_Imie1_Nazwisko2_Imie2_PD3.zip`. W archiwum znajdować się powinien jeden katalog, w którym umieszczone zostaną następujące pliki:

- prezentację (slajdy) przedstawiającą wyniki analizy danych (PDF lub HTML) – to *głównie* na jej podstawie zostanie wystawiona ocena;
- wszystkie skrypty `.R`, moduły / skrypty `.py`, notatniki pozwalające na odtworzenie zawartych w prezentacji wyników.

Prezentacje: Zgodnie z terminami wskazanymi w harmonogramie przedmiotu na XIV i /lub XV zajęciach każda grupa przedstawi najciekawsze ich zdaniem wyniki. Prezentacja grupy powinna trwac nie dłużej niż **10 minut na prezentację projektu dla zespołów 2-osobowych, 15 dla 3-osobowych**. Wygłoszenie prezentacji jest warunkiem koniecznym uzyskania pozytywnej oceny.

($\star$) W przypadku gdy zespół składa się z osób z różnych grup laboratoryjnych prowadzący wskazuje w której grupie zajęciowej zespół wygłosi prezentację.

1 Dane do analizy

Będziemy pracować na danych dotyczących wydajności czasowej linii lotniczych, które zostały przygotowane i upublicznione w ramach konkursu *American Statistical Association Data Expo 2009:* **Airline on-time performance**. Aktualnie, dane te są dostępne w Repozytorium *Harvard Dataverse* (por. [2008, "Data Expo 2009: Airline on time data", https://doi.org/10.7910/DVN/HG7NV7, Harvard Dataverse, V1]).

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/HG7NV7

Dane składają się ze szczegółów przylotów i odlotów wszystkich lotów komercyjnych na terenie USA, od października 1987 do kwietnia 2008. Wśród zawartych w nich informacjach znajdziemy m.in.:

- `Year`: rok (1987-2008);
- `Month`: miesiąć (1-12);
- `DayofMonth`: dzień miesiąca (1-31);
- `DayOfWeek`: dzień tygodnia (1-7; Monday - Sunday);
- `DepTime`: czas wylotu (local, hhmm);
- `CRSDepTime`: planowany czas wylotu (local, hhmm);
- `ArrTime`: aktualny czas przylotu (local, hhmm);
- `CRSArrTime`: planowany czas przylotu (local, hhmm);
- `UniqueCarrier`: unikalny kod przewoźnika;
- `FlightNum`: numer lotu;
- `TailNum`: identyfikator samolotu;
- `ActualElapsedTime`: czas całkowity (w minutach);
- `CRSElapsedTime`: planowany czas (w minutach);
- `AirTime`: czas "w powietrzu" (w minutach);

[1] A więc nie: .rar, .7z itp.

- `ArrDelay`: opóźnienie przylotu (w minutach);
- `DepDelay`: opóźnienie wylotu (w minutach);
- `Origin`: kod lotniska wylotu;
- `Dest`: kod lotniska przylotu;
- `Distance`: odległość w milach;
- `Cancelled`: czy lot został odwołany?
- `CancellationCode`: powód odwołania lotu (A = przewoźnik, B = pogoda, C = NAS, D = bezpieczeństwo).

2 Cel projektu

Niniejsza praca domowa to prawdziwe wyzwanie *data science* – to każda grupa sama stawia ciekawe (dla siebie i słuchaczy) pytania i generuje na nie odpowiedzi. Interesują nas odpowiedzi na pytania dotyczące zarówno czynników mających wpływ na opóźnienia, jak i ogólnie zależności w czasie (np. zmiany w natężeniu ruchu powietrznego, liczby pasażerów itp.). Poniżej znajdą Państwo kilka przykładowych pytań:

1. Kiedy jest najlepsza pora dnia/dzień tygodnia/pora roku na lot, aby zminimalizować opóźnienia?
2. Czy starsze samoloty mają więcej opóźnień?
3. Jak zmienia się liczba osób latających pomiędzy różnymi lokalizacjami w czasie?
4. Jak bardzo pogoda wpływa opóźnienia samolotów?
5. Czy można wykryć awarie kaskadowe, gdy opóźnienia na jednym lotnisku powodują opóźnienia na innych?

3 Ocena

Ocenę co najmniej dostateczną (> 50% - min. 15 pkt) uzyskają prace, które spełniają następujące kryteria:

1. zawierają kody potrzebny do wczytania zbiorów danych oraz generowania wszystkich zawartych w prezentacji wyników (tzn. tabel, wykresów);
2. stworzą kod, dzięki któremu zostaną wygenerowane co najmniej **dwa** dla zespołów dwuosobowych / **trzy** dla zespołów trzyosobowych ciekawe wyniki (odpowiedzi na pytania „badawcze'' w postaci wykresów/tabel/itp.);
3. przedstawią uzyskane wyniki w formie prezentacji (10 min zespoły 2-osobowe /15 min zespoły trzyosobowe).

Każda dodatkowa analiza, wykres, ciekawa zastosowana technika będzie wpływać pozytywnie na ocenę (np. wykresy interaktywne, aplikacja w *Shiny*, animacje, nietrywialność stawianych pytań, korzystanie ze źródeł zewnętrznych np. w celu wyjaśnienia odkrytych zależności).

W szczególności, ocenę maksymalną (bardzo dobrą) uzyskają prace wyróżniające się pod względem jakościowym oraz merytorycznym.

Pomocnicze zbiory danych (źródła zewnętrzne), z których możemy korzystać:

- http://ourairports.com/data/
- http://openflights.org/data.html
- http://opensky-network.org/