PD-R-Py 2019/2020

Praca projektowa nr 3 (max. = 30 p.)

Zadanie rozwiązujemy w grupach dwuosobowych.

Termin oddania pracy: ~~02.06.2020, godz. 06:00~~ **08.06.2020, godz. 06:00**

Do przesłania na adres `A.Cena@mini.pw.edu.pl` ~~prowadzącego laboratoria `M.Bartoszuk@mini.pw.edu.pl` lub `A.Cena@mini.pw.edu.pl`~~ ze swojego konta pocztowego `*@*pw.edu.pl` **jeden** plik `.zip` (nie: `.rar`, `.7z` itp.) o nazwie typu

`Nazwisko1_Imie1_NrAlbumu1_Nazwisko2_Imie2_NrAlbumu2_pd3.zip`, zawierający:

- prezentację (slajdy) zawierającą omówienie sposobu rozwiązania zadania oraz przedstawiającą wyniki analizy danych (PDF lub HTML) – to *głównie* na jej podstawie zostanie wystawiona ocena;
- wszystkie skrypty `.R` i notatniki pozwalające na odtworzenie zawartych w prezentacji wyników;
- dane pośrednie, na podstawie których zostały wygenerowane wyniki (pliki `.csv`, `.json` itp.); uwaga: *nie* dodajemy plików zawierających dane surowe – przesyłany plik `.zip` powinien być „rozsądnych" rozmiarów;
- zdjęcie lub skan oświadczenia o samodzielności wykonanej pracy (treść znajduje się w sekcji `Oświadczenie`) od każdego z członków zespołu.

Uwaga: Lepiej by nazwy plików nie zawierały polskich liter diakrytyzowanych (przekształć $ą \to a$ itd.).

Prezentacje: Na XIV i XV zajęciach każda dwuosobowa grupa przedstawi najciekawsze ich zdaniem wyniki (10 minut na prezentację projektu + 5 minut na dyskusję i pytania od słuchaczy). Wygłoszenie prezentacji jest warunkiem koniecznym uzyskania pozytywnej oceny.($\star$)

($\star$) Jeśli zajęcia na uczelni nie zostaną odwieszone sposób „wygłoszenia" wygłoszone prezentacje przesyłamy w formie filmu nagranego i zmontowny w ramach zespołu - np. każdy uczestnik narywa część dotyczącą swojego wystąpienia - film udostępniają Państwo w ramach kanału `youtube` jako `unlisted` (wtedy będzie dostępy tylko dla osób posiadających link) i przesyłają Państwo link prowadzącym (film będzie udostępniony także innym uczestnikom grupy).

W ramach MS Teams odbędzie się dyskusja na temat wygłoszonych/odsłuchanych prezentacji w godzinach zajęć w terminie ustalonym przez prowadzącego.

($\star$) W przypadku gdy zespół składa się z osób z różnych grup laboratoryjnych prowadzący wybiera termin wygłoszenia prezentacji / uczestnictwa w dykusji.

1 Dane do analizy

Będziemy pracować na danych udostępnionych przez NYC Bike Share, LLC oraz Jersey City Bike Share, LLC dotyczących klientów korzystających z systemu rowerów miejskich oraz ich podróży. Dane zawierają następujące zmienne:

1. Trip Duration (seconds)
2. Start Time and Date

3. Stop Time and Date
4. Start Station Name
5. End Station Name
6. Station ID
7. Station Lat/Long
8. Bike ID
9. User Type (Customer = 24-hour pass or 3-day pass user; Subscriber = Annual Member)
10. Gender (Zero=unknown; 1=male; 2=female)
11. Year of Birth

Dane są dostępne do pobrania na stronie[1]:

https://www.citibikenyc.com/system-data

Interesują nas zbiory o nazwach w formacie `rrrrmm-citibike-tripdata.csv.zip` - dotyczące Nowego Jorku, np.
`201909-citibike-tripdata.csv.zip` (dane września 2019 r.) lub dane

`JC-rrrrmm-citibike-tripdata.csv.zip` (dotyczące Jersey City).

Pliki są różnych rozmiarów np. zawierające dane dotyczące Nowego Jorku są z reguły większe niż pliki dotyczące Jersey City. W obu przypadkach struktura zbioru powinna być taka sama (te same kolumny i ich nazwy, format daty itd.). W obu przypadkach należy dane przetworzyć np. zaagregować, wybrać / przefiltrwać interesujące nas informacje i/lub zapisać je w mniejszym pliku.

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 pytania zarówno na temat użytkowaników jak i odbytych przez nich podróży, np. w którym dniu tygodnia rowerem jeździmy najdłużej? Z której stacji najczęściej? W jakich miesiącach ruch rowerowy jest duży? Czy zmienia się to w czasie? itp.

Do analizy należy wykorzystać dane dotyczące **co najmniej sześciu miesięcy**, przy czym nie muszą to być kolejne miesiące lub nawet lata. Im więcej danych Państwo wykorzystają tym ciekawsze można uzyskać wyniki np. pokazać jakieś zależności w czasie.

2 Ocena

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

1. zawierają kody potrzebny do generowania wyników w tym m.in. wczytywania zbiorów,
2. stworzą kod, dzięki któremu zostaną wygenerowane co najmniej dwa ciekawe wyniki (odpowiedzi na pytania „badawcze" w postaci wykresów/tabel/itp.),
3. przedstawią uzyskane wyniki w formie prezentacji.

Każda dodatkowa analiza, wykres, ciekawa zastosowana technika będzie wpływać pozytywnie na ocenę (np. wykresy interaktywne, aplikacja w *Shiny*, animacje, algorytmy i struktury danych umożliwiające poprawę szybkości wykonywanych analiz, własne implementacje metod znanych z literatury (z autorskimi modyfikacjami) itp.), nietrywialność stawianych pytań itd.

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

[1] Proszę zapoznać się z informacjami dotyczącymi danych udostępnionymi na stronie w tym z licencją ich użytkowania https://www.citibikenyc.com/data-sharing-policy

3 Oświadczenie

Do przesłanej pracy każdy z członków zespołu dołącza własnoręcznie podpisane oświadczenie o następującej treści:

Oświadcznie w sprawie pracy projektowej nr 3

Oświadczam, że niniejsza praca stanowiąca podstawę do uznania efektów uczenia się z przedmiotu

Przetwarzanie danych w języku R i Python

została wykonana samodzielnie w ramach zgłoszonego zespołu.

Imię i Nazwisko

Nr albumu