Kod szkolenia: **R/DPLYR**

Tytuł szkolenia: **Nowoczesne przetwarzanie danych w R**

Dni: 3

Opis:

Adresaci szkolenia:

Szkolenie jest adresowane do szerokiego grona użytkowników języka R (analityków danych, badaczy danych, statystyków oraz programistów), którzy zainteresowani są opanowaniem umiejętności szybkiego tworzenia przejrzystego, elastycznego oraz łatwego w utrzymaniu kodu służącego wstępnemu zrozumieniu danych oraz przygotowaniu ich na potrzeby modelowania. Szkolenie objaśnia zarówno podstawowe, jak też zaawansowane aspekty wykorzystania pakietów dplyr oraz purrr. Na udziale w nim mogą skorzystać zarówno początkujący, jak też bardziej zaawansowani użytkownicy języka R.

Cel szkolenia:

Uczestnicy szkolenia opanują umiejętność wstępnej analizy danych oraz przetwarzania danych w języku R z wykorzystaniem pakietów dplyr oraz purrr pracując nad praktycznymi zagadnieniami między innymi z obszaru inżynierii cech (feature engineering), selekcji cech (feature selection), czyszczenia danych (data cleaning) oraz analiz mających na celu wstępne zrozumienie danych.

W szczególności:

- będą sprawnie posługiwać się najużyteczniejszymi w codziennej pracy funkcjonalnościami tych pakietów,
- zdobędą wiedzę o szerokiej gamie dostępnych w nich funkcji, które pozwalają na rozwiązywanie mniej standardowych problemów,
- poznają zaawansowane aspekty pracy z pakietem dplyr w tym współpracy pakietu z bazą danych,
- uzyskają ogólną wiedzę na temat możliwości pakietów komplementarnych oraz alternatywnych.

Mocne strony szkolenia:

Szkolenie zapozna uczestnika z najlepszymi praktykami przetwarzania danych w języku R ucząc:

unikalnej składni wykorzystywanej przez pakiet dplyr do przejrzystego tworzenia kodu analitycznego,

pełnego spektrum funkcjonalności pakietu dplyr dającego swobodę w wyborze optymalnego rozwiązania wielu problemów,

najciekawszych z punktu widzenia przetwarzania danych aspektów pakietu purrr.

Nauka prowadzona będzie w oparciu o przykłady oraz zadania warsztatowe zaczerpnięte z praktyki pracy statystycznej na etapie poznawania danych oraz przygotowania ich na potrzeby modelowania.

Wymagania:

Efektywne skorzystanie ze szkolenia wymaga posiadania podstawowej wiedzy w zakresie programowania w R. W szczególności przydatna będzie znajomość podstawowej składni języka (instrukcja warunkowa if, pętla for, umiejętność tworzenia własnych prostych funkcji), znajomość podstawowych struktur danych (wektor, lista, ramka danych) oraz umiejętność dokonywania podstawowych operacji na danych jak wyznaczenie wartości średniej z wektora. Szkolenie realizowane jest z wykorzystaniem bazy danych MySQL do której komputery szkoleniowe muszą mieć dostęp.

Parametry szkolenia:

3 * 8 godzin (3 * 7 neto). Przewaga warsztatów skupionych na pracy z danymi.

Wielkość grupy maksymalnie 8-10 osób.

Program szkolenia:

1. Szybki wstęp do tematyki przetwarzania danych.
2. Obiekt tibble jako narzędzie wygodnej pracy z danymi w konsoli R.
3. Wprowadzenie do mechaniki przetwarzania danych z pakietem dplyr.
4. Zapoznanie z podstawowymi funkcjonalnościami pakietu dplyr.
    - Manipulowanie zmiennymi.
    - Manipulowanie obserwacjami.
    - Analizowanie danych.
5. Podnoszenie komfortu pracy z pakietem dplyr.
    - Zwinne metody selekcji kolumn.

- Tworzenie i modyfikowanie zmiennych.
- Wybieranie i porządkowanie obserwacji.
- Dodatkowe funkcjonalności pakietu dplyr.

6. Masowe przetwarzanie zmiennych.
   - Równoczesne przetwarzanie wszystkich zmiennych.
   - Równoczesne przetwarzanie wybranych zmiennych.
   - Równoczesne warunkowe przetwarzanie zmiennych.
7. Funkcje okienkowe z pakietem dplyr.
   - Zapoznanie z podstawową konstrukcją programistyczną.
   - Przegląd różnych możliwości zastosowania.
8. Łączenie zbiorów danych.
   - Proste łączenie wierszy lub kolumn.
   - Łącznie z wykorzystaniem kolumn kluczy.
9. Szybkie wprowadzenie do pakietu purrr.
10. Transformowanie zbiorów danych z wykorzystaniem pakietu purrr.
    - Warunkowa selekcja kolumn.
    - Złożone transformacje kolumn z użyciem map().
    - Warunkowe funkcje map().
    - Użyteczne rozszerzenia funkcji map().
    - Kumulowanie wyników pośrednich.
11. Strategia split-apply-combine.
    - Wprowadzenie do strategii z wykorzystaniem R base.
    - Prosta implementacja strategii z wykorzystaniem pakietu dplyr.
    - Zaawansowane wykorzystanie strategii w oparciu o pakiet purrr.
    - Funkcje wspierające pracę z wynikami cząstkowymi.
12. Zaawansowane aspekty pracy z pakietem dplyr.
    - Elastyczne przetwarzanie danych z funkcją do().
    - Uzupełnienie wiedzy na temat łączenia zbiorów danych.
    - Zaawansowane aspekty składni pakietu dplyr.
13. Współpraca pakietu dplyr z bazą danych.
    - Wykorzystanie pakietu dbplyr do łączenia z bazą danych.
    - Funkcje wspierające pracę z bazami danych w pakiecie dplyr.
14. Szybki przegląd rozwiązań komplementarnych i alternatywnych.
    - Czyszczenie danych z pakietem tidyr.
    - Zmiana struktury danych z pakietem reshape.
    - Pakiet data.table jako alternatywa dla dplyr.