

Metody czyszczenia danych  
Karta opisu przedmiotu (sylabus)

## Informacje podstawowe

<b>Kierunek studiów</b> Informatyka i analityka danych <b>Specjalność</b> Analityka Gospodarcza <b>Jednostka organizacyjna</b> UEP <b>Poziom kształcenia</b> studia pierwszego stopnia (licencjackie) <b>Forma studiów</b> stacjonarne <b>Profil kształcenia</b> ogólnoakademicki		<b>Cykl dydaktyczny</b> 2025/2026 <b>Kod przedmiotu</b> UEPIADAGS.14C.205933.25 <b>Język wykładowy</b> Polski <b>Obligatoryjność</b> Obowiązkowy <b>Blok zajęciowy</b> Blok C		
<b>Osoba odpowiedzialna za treść sylabusu</b>		Maciej Beręsewicz		
<b>Okres</b> Semestr 3		<b>Forma zaliczenia</b> Zaliczenie  <b>Forma prowadzenia i godziny zajęć</b> • Uczestnictwo w wykładach: 30		<b>Liczba punktów ECTS</b> 2

## Cele uczenia się dla przedmiotu

C1	Wykształcenie umiejętności identyfikacji problemów związanych z pracą na zbiorach danych rzeczywistych
C2	Wykształcenie umiejętności identyfikacji różnych źródeł danych i sposobów ich importowania w celu dalszej analizy
C3	Uzyskanie wiedzy z zakresu techniki walidacji i czyszczenia danych celem przygotowania ich do analizy
C4	Uzyskanie wiedzy z zakresu pakietów języka R służących do walidacji i czyszczenia danych

## Cele kształcenia UEP

Kod	Treść celu
CS1_1.1	student rozumie strukturę i elementy problemu gospodarczego
CS1_1.2	student właściwie wykorzystuje techniki, metody i modele analityczne
CS1_2.1	student rozumie i stosuje zasady pracy grupowej
CS1_2.2	student skutecznie komunikuje się w formie ustnej i pisemnej
CS1_3.2	student stosuje etyczne standardy akademickie i biznesowe, uwzględniając aspekty środowiskowe i społeczne

## Wymagania wstępne

Podstawowa znajomość R i statystyki.

## Efekty uczenia się dla przedmiotu

Kod	Efekty w zakresie	Kierunkowe efekty uczenia się	Metody weryfikacji
<b>Wiedzy</b>			
W1	Student rozpoznaje problemy związane z pracą na zbiorach danych rzeczywistych	K1_W04, K1_W05	Projekt grupowy / praca w grupie
W2	Student rozpoznaje techniki walidacji i czyszczenia danych mające na celu przygotowanie ich do analizy	K1_W04, K1_W05	Projekt grupowy / praca w grupie
W3	Student rozpoznaje pakiety języka R służące do walidacji i czyszczenia danych	K1_W04, K1_W05	Projekt grupowy / praca w grupie
<b>Umiejętności</b>			
U1	Student stosuje główne źródła danych i wie, jak je pozyskać do dalszej analizy.	K1_U01, K1_U06	Projekt grupowy / praca w grupie
U2	Student wczytuje dane zapisane w różnych formatach	K1_U01, K1_U05	Projekt grupowy / praca w grupie
U3	Student dokonuje walidacji zbioru danych w kontekście wykorzystania ich do analizy	K1_U01, K1_U02	Projekt grupowy / praca w grupie
U4	Student czyści dane w celu przygotowania ich do analizy	K1_U01, K1_U02	Projekt grupowy / praca w grupie
<b>Kompetencji społecznych</b>			
K1	Student ma świadomość problemów związanych z pracą na zbiorach danych rzeczywistych	K1_K01	Projekt grupowy / praca w grupie
K2	Student jest gotów do rzetelnej i odpowiedzialnej etycznie pracy na zbiorach danych rzeczywistych	K1_K01, K1_K04, K1_K05	Projekt grupowy / praca w grupie
K3	Jest zorientowany na uzupełnianie wiedzy w zakresie narzędzi informatycznych służącym do pracy na zbiorach danych	K1_K01	Projekt grupowy / praca w grupie

## Treści programowe

Lp.	Treści programowe	Cele kształcenia dla przedmiotu	Efekty uczenia się dla przedmiotu
-----	-------------------	---------------------------------	-----------------------------------

Lp.	Treści programowe	Cele kształcenia dla przedmiotu	Efekty uczenia się dla przedmiotu
1.	Reprezentacja danych w pakietach statystycznych	C2, C3	W1, W3, U1, U4, K1, K2, K3
2.	Dostęp do najistotniejszych publicznych źródeł danych w Europie i główne pakiety R do ich importowania.	C1, C2	W1, U1, U2, K1
3.	Wczytywanie danych z plików o różnym formacie (m.in. csv, xlsx, xml, json) oraz baz danych	C1, C2, C3	W1, W3, U1, U2, K1, K3
4.	Przekształcanie zbiorów danych (układ wąski i szeroki)	C1, C2, C3	W1, W2, W3, U2, U3, U4, K1, K2, K3
5.	Weryfikacja i walidacja zbiorów danych	C1, C2, C3, C4	W1, W2, W3, U2, U3, U4, K1, K2, K3
6.	Przetwarzanie danych tekstowych: wyrażenia regularne, powiązane pakiety i modele sztucznej inteligencji do podsumowywania danych tekstowych	C1, C3	W1, W3, U3, U4, K1, K2, K3
7.	Operacje na datach	C1, C2, C3	W1, W2, W3, U3, U4, K1, K2, K3
8.	Podstawy automatycznej edycji danych i imputacji dedukcyjnej	C1, C2, C3, C4	W1, W2, W3, U3, U4, K1, K2, K3

## Literatura

### Obowiązkowa

1. van der Loo, M., & de Jonge, E. (2018). Statistical data cleaning with applications in R. John Wiley & Sons.
2. Gagolewski, M. (2022). stringi: Fast and portable character string processing in R. Journal of Statistical Software, 103, 1-59.
3. Wickham, H. . (2014). Tidy Data. Journal of Statistical Software, 59(10), 1-23. <https://doi.org/10.18637/jss.v059.i10>
4. Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. The American Statistician, 72(1), 2-10.
5. Golemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. Journal of statistical software, 40, 1-25.

### Zalecana

1. Wickham H., Golemund G., and Çetinkaya-Rundel, M. (2023), R for Data Science (2e). <https://r4ds.hadley.nz/>
2. De Waal, T., Pannekoek, J., & Scholtus, S. (2011). Handbook of statistical data editing and imputation (Vol. 563). John Wiley & Sons.
3. Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. " O'Reilly Media, Inc.". <https://www.tidytextmining.com/>

## Informacje rozszerzone

### Metody nauczania:

Metoda projektów , Wykład konwersatoryjny, Wykład z prezentacją multimedialną, Dyskusja, Analiza przypadków, Rozwiązywanie zadań, Ćwiczenia laboratoryjne, Metody e-learningowe, Praca z komputerem

Metody nauczania	Sposób zaliczenia	Warunki zaliczenia przedmiotu
Wykład	Projekt grupowy / praca w grupie	Przynajmniej 50% z projektu

## Rozliczenie punktów ECTS

<b>Forma aktywności studenta</b>	<b>Średnia liczba godzin* przeznaczonych na zrealizowane aktywności</b>	
Uczestnictwo w wykładach	30	
Przygotowanie projektu	25	
Konsultacje z prowadzącym/i zajęcia	5	
<b>Łączny nakład pracy studenta</b>	<b>Liczba godzin</b> 60	<b>ECTS</b> 2.0
<b>Zajęcia z bezpośrednim udziałem nauczyciela</b>	<b>Liczba godzin</b> 35	<b>ECTS</b> 1.0
<b>Nakład pracy związany z zajęciami o charakterze praktycznym</b>	<b>Liczba godzin</b> 25	<b>ECTS</b> 1.0

\* godzina (lekcyjna) oznacza 45 minut

## Kierunkowe efekty uczenia się

Kod	Treść
K1_K01	absolwent jest gotów do obiektywnej, krytycznej oceny swojej wiedzy i umiejętności
K1_K04	absolwent rozumie społeczne, prawne i etyczne aspekty pracy z danymi, w tym związane z ochroną prywatności i bezpieczeństwem informacji
K1_K05	absolwent jest gotów do kierowania się uczciwością intelektualną w działaniach własnych i wymagania jej od innych osób, jest wyczulony na kwestie plagiatu, auto-plagiatu i wykorzystania sztucznej inteligencji z poszanowaniem zasad etyki
K1_U01	absolwent potrafi poprawnie dobrać i efektywnie wykorzystywać metody i narzędzia ilościowe i informatyczne do opisu i analizy zjawisk społeczno-gospodarczych
K1_U02	absolwent potrafi właściwie dobierać źródła informacji, dokonywać ich oceny oraz przeprowadzać krytyczną analizę i syntezę danych z różnych źródeł, w celu wyciągania trafnych wniosków i podejmowania optymalnych decyzji w warunkach nie w pełni przewidywalnych
K1_U05	absolwent potrafi planować i organizować pracę indywidualną oraz zespołową zorientowaną na rozwiązywanie specjalistycznych problemów
K1_U06	absolwent jest gotów do systematycznego podnoszenia swoich kompetencji poprzez ciągłe samokształcenie w oparciu o zaawansowaną literaturę specjalistyczną
K1_W04	absolwent zna i rozumie podstawy prawa, ochrony własności i prawa autorskiego, a także prawa i etyki w zakresie stosowania nowoczesnych technologii informacyjnych z uwzględnieniem wykorzystania sztucznej inteligencji
K1_W05	absolwent zna i rozumie zasady tworzenia i rozwoju różnych form przedsiębiorczości, w szczególności opartych na wykorzystaniu nowoczesnych technologii informacyjno-komunikacyjnych

Ten utwór jest dostępny na licencji Creative Commons Uznanie autorstwa 4.0 Międzynarodowa (CC BY 4.0). Autor: Maciej Beręsewicz