

Ćwiczenia 1.

Informacje ogólne dotyczące ćwiczeń.

Wprowadzenie do R i RStudio

1. Organizacja zajęć

- Udział w ćwiczeniach jest obowiązkowy.
- Ćwiczenia odbywają się hybrydowo- 6 lub 7 spotkań on-line platforma MS Teams, 1-2 (4-6h) spotkania stacjonarne (warsztaty z modelowania danych rzeczywistych).
- Możecie mieć Państwo 1 nieobecność nieusprawiedliwioną.
- Każdą kolejną nieobecność należy odrobić- w postaci sprawozdania z zajęć zawierającego analizę, interpretację wyników, wnioski. Dodatkowo sprawozdanie powinno zawierać uzasadnienie wyboru zmiennych w modelu, parametrów modelu. Sprawozdanie należy oddać najpóźniej na 1 zajęciach po ustaniu nieobecności.
- Posiadanie 50% lub większej liczby nieobecności= brak zaliczenia z ćwiczeń.
- W trakcie semestru mogą odbywać się kartkówki sprawdzające przygotowanie do zajęć.
- W trakcie semestru zaplanowane jest 1 zadanie domowe:
 - Zadanie z preprocessingu- wykonywane w parach, lub
 - Tutorial wykonania modelu data miningowego- wykonywany w grupach 4 osobowych.
- Pod koniec semestru dostaniecie Państwo projekty zaliczeniowe Case Study (wylosowane dane i metody). Są 3 terminy oddania projektu zaliczeniowego (koniec zajęć, koniec sesji podstawowej, koniec sesji poprawkowej).
- Aby uzyskać zaliczenie z ćwiczeń należy zdobyć minimum 50% punktów możliwych do uzyskania w semestrze, przy jednoczesnym posiadaniu minimum 50% punktów z Case Study .
- Ocena końcowa z ćwiczeń: 80% CS+ 15% Zadanie/Tutorial+5% Kartkówka.
- Termin oceny końcowej jest związany z terminem spełnienia wszystkich wymagań dotyczących zaliczenia ćwiczeń.
- Ocena końcowa jest oceną zgodną z Regulaminem Studiów AGH.
- Osoby aktywne na zajęciach mogą być dodatkowo punktowane.
- Ocena końcowa z zajęć jest oceną końcową z ćwiczeń.
- Materiały do ćwiczeń dostępne są na stronie [www: monikachuchro.pl](http://www.monikachuchro.pl) oraz na kanale ogólnym w MS Teams.
- Jeśli jakieś metody, algorytmy się powtarzają, proszę o informację.
- Dodatkowe informacje są zawarte w sylabusie przedmiotu.
- W przypadku pytań ogólnych dotyczących zajęć proszę je umieszczać na kanale ogólnym zajęć. Możecie się też Państwo kontaktować przez wiadomość e-mail: chuchro@agh.edu.pl (proszę o nie wysyłanie wiadomości na czacie teamsów).
- Starostę każdej z grup oraz starostę roku proszę o kontakt.
- Materiały dodatkowe do zajęć (polecam Eksploracja danych D. Hand, dokumentacje techniczne, <https://journal.r-project.org/>).

2. Na 2 zajęcia proszę powtórzyć:

- tworzenie zmiennych, ekstrakcja danych, zapisywanie: wektorów, faktorów, dataframe,
- tworzenie własnych funkcji,
- instrukcje warunkowe, pętle,
- praca z NA,
- import, eksport danych,
- tworzenie wizualizacji: wykresy liniowe, box-plot, rozrzutu, histogramy,
- interpretacja statystyk opisowych, wykresów statystycznych, macierzy korelacji,
- raport w RMarkdown.

3. Oprogramowanie

Proszę zainstalować lub uaktualnić R i RStudio do najnowszej wersji oraz uaktualnić pakiety tidyverse.

Powtórzenie materiału z Programowania w R i wizualizacji danych- sprawdź czy umiesz!

Zadanie 1. Wektor

- Utwórz zmienną, która będzie przechowywać liczbę osób, które uczęszczają na zajęcia w danym dniu tygodnia. Nazwij zmienną *lista_obecnosci* i nadaj nazwy poszczególnym wartościom.
- Wartości zmiennej: 25, 30, 16, 20, 10, 21, 12.
- Wyświetl, ile osób uczestniczyło w zajęciach w poniedziałek i piątek.
- Wyświetl dni z największą i najmniejszą liczbą uczestników.
- Oblicz, ile średnio osób uczestniczyło w zajęciach.
- Policz w ile dni frekwencja była wyższa lub równa 20 osób.
- Utwórz zmienną, która przechowuje liczbę osób, która powinna brać udział w zajęciach. Nazwij ją: *liczebosc_grup* i przypisz jej wartości: 25, 30, 18, 20, 15, 21, 15.
- Sprawdź w jakie dni na zajęciach uczestniczyło 100% osób.
- Sprawdź w jakie dni na zajęciach uczestniczyło mniej niż 50% osób.

Zadanie 2. Macierz

- Napisz kod, który da następujący wynik:

	<i>US</i>	<i>non-US</i>
<i>A New Hope</i>	460.998	314.4
<i>The Empire Strikes Back</i>	290.475	247.9
<i>Return of the Jedi</i>	309.306	165.8

- Wybierz z macierzy ostatni wiersz danych.
- Wybierz z macierzy wartość 314.4.

Zadanie 3. Dataframe

- Napisz kod, który da następujący wynik:

	name	type	diameter	rotation	rings
1	Mercury	Terrestrial planet	0.382	58.64	FALSE
2	Venus	Terrestrial planet	0.949	-243.02	FALSE
3	Earth	Terrestrial planet	1.000	1.00	FALSE
4	Mars	Terrestrial planet	0.532	1.03	FALSE
5	Jupiter	Gas giant	11.209	0.41	TRUE
6	Saturn	Gas giant	9.449	0.43	TRUE
7	Uranus	Gas giant	4.007	-0.72	TRUE
8	Neptune	Gas giant	3.883	0.67	TRUE

- Wybierz średnice dla Marsa (diameter),
- Wybierz całość danych dla Uranus,
- Sprawdź strukturę utworzonej data frame
- Wyświetl informacje o pierścieniach (rings),
- Utwórz zmienną planets_rings w której będą tylko te planety które mają pierścienie,
- Utwórz podzbiór zawierający planety o średnicy mniejszej niż 1,

Zadanie 4. Wektory czynnikowe

- Utwórz 2 wektory czynnikowe (factors), w tym jeden typu ordered.
- Sprawdź typ danych.
- Policz liczebność obserwacji należących do określonej klasy.

Zadanie 5. Funkcja

Napisz funkcję, która będzie wystawiać oceny na podstawie arkusza i Regulaminu Studiów AGH. Ocena może być wyliczona jeśli Student ma minimum 50% obecności. Jeśli nie spełnia tego warunku, zamiast oceny powinno wyświetlać się „brak zaliczenia”. Waga dla kolokwium to 0.9, dla każdej z pozostałych aktywności 0.05. W przypadku nie zaliczenia kolokwium ocena końcowa 2.0.

	liczba obecności	ocena kolokwium	ocena kartkowka	ocena aktywnosc
max	15	100	100	100

Proszę przetestować poprawność działania funkcji.

Zadanie 6. Pakiet dplyr

- Wykonaj filtrowanie danych gapminder dla Chin z roku 2002.
- Wykonaj filtrowanie danych gapminder dla roku 1957 i wykonaj sortowanie wyników według zmiennej pop.
- Dla danych z roku 2007, utwórz nową zmienną w obrębie gapminder
 - $\text{lifeExpMonths} = \text{lifeExp} * 12$
 - posortuj wyniki według nowo utworzonej zmiennej od najwyższych do najniższych wartości.
- Dla danych z roku 1957 oblicz medianę z lifeExp i umieść w zmiennej medianLifeExp oraz maksimum z gdpPercap i umieść w zmiennej maxGdpPercap.
- Dla pogrupowanych danych według kontynentu i roku, oblicz medianę z lifeExp i umieść w zmiennej medianLifeExp oraz wartość średnią z gdpPercap i umieść w zmiennej meanGdpPercap.

Zadanie 7. Import i export danych

- Zaimportuj pliki za pomocą funkcji i przypisz do odpowiadających im nazw zmiennych,
- Sprawdź poprawność danych,
- Wyeksportuj zmienne do pliku *.csv,
- Powtórz z wykorzystaniem RStudio okna importu.

Zadanie 8.

Dla zaimportowanych danych z zadania 7 wykonaj odpowiednie wykresy wykorzystując pakiet graphics i ggplot2. Wykresy do wykonania:

- histogram,
- wykres rozrzutu,
- wykres ramka-wąsy z uwzględnieniem zmiennej jakościowej,
- wykres liniowy.

Zadanie 9.

Utwórz raport RMarkdown z zadań 2, 3, 5, 8. Wygeneruj go do wybranego typu dokumentu. W dokumencie powinny być widoczne fragmenty kodu, wyniki, wizualizacje oraz Państwa komentarze.