LABORATORIUM 6.

ZARZĄDZANIE DANYMI

RAMKA DANYCH:

wywołanie kolumny w ramce np. Ankieta$Płeć
usuwanie kolumny w ramce np. Ankieta[,–nr] lub Ankieta$Płeć = NULL
edycja ramki np. *fix*(Ankieta)
symbol braku danych – NA

WYKRES:

histogram dla zmiennej kategorycznej – *ggplot* (zbiór danych, *aes* (*x* = zmienna)) + *geom_bar* (*fill* = ”kolor”, *col* = ”kolor”) + *ylab* (”opis”)

ZAD.1. Otworzyć plik z danymi w Excel

a) zmienić nagłówki kolumn na krótsze wprowadzając nazwy: Waga, Wzrost, M.zamieszkania, Sz.średnia, ECTS, Algebra, MSzS1, Narz.inż, Prog1, WdI, L.godzin, L.sys.op, System, Wiek;

b) w kolumnie Sz.średnia zamienić łańcuchy *klasa z rozszerzoną matematyką* na *RM*;

c) plik po zmianach zapisać pod nazwą Ankieta.xlsx.

ZAD.2. Wczytać w RStudio plik Ankieta.xlsx (Environment/Import Dataset/From Excel), zmieniając na *numeric* typ zmiennych mierzalnych: Waga, Wzrost, ECTS, Algebra, MSzS1, Narz.inż, Prog1, WdI, L.godzin, L.sys.op, Wiek

a) wyświetlić podsumowanie danych (użyć *summary*) przed i po faktoryzacji zmiennych niemierzalnych (użyć *factor*) – ocenić wartości skrajne zmiennych mierzalnych;

b) w ramce Ankieta utworzyć nową zmienną Średnia, zawierającą średnią ocen z kursów;

c) przenieść kolumny z ocenami z kursów do podzbioru Ankieta.kursy (użyć *subset*);

d) napisać funkcję **zakres3sigm**, która zwróci dla dowolnej zmiennej ramkę danych z nagłówkami *lewy.kres* / *prawy.kres* jako średnią minus / plus trzy odchylenia standardowe (użyć *function*, *mean*, *sd*, *data.frame*);

e) dla zmiennej Średnia wyznaczyć ewentualne dane odstające i zastąpić je symbolem braku danych, a później średnią, zaokrągloną do części dziesiętnych (przy dużej liczbie danych użyć *which*);

f) utworzyć podzbiory danych Ankieta.M i Ankieta.K dla mężczyzn i kobiet odpowiednio (użyć *filter*);

g) dla zmiennych Waga i Wzrost wyznaczyć ewentualne dane odstające dla obu płci, zastąpić je symbolem braku danych, a później średnią, zaokrągloną do części dziesiątych;

h) utworzyć nową zmienną L.g.kody, w której zostaną umieszczone 3 przedziały liczbowe odpowiadające ustalonym kategoriom: krótko, średnio, długo (użyć *cut*) i wyświetlić liczności przedziałów.

ZAD.3. Wyznaczyć histogramy dla zmiennych M.zamieszkania, Sz.średnia i System.