Ćwiczenia 2.

Preprocessing danych

Problemy występujące w danych:

- braki danych,

- błędne wartości (błędy grube- sprzeczne z logiką),

-wartości outliers (ekstremalne i odstające),

-duplikaty,

-przestarzałe atrybuty, wiarygodność danych, brak ujednolicenia jednostek…

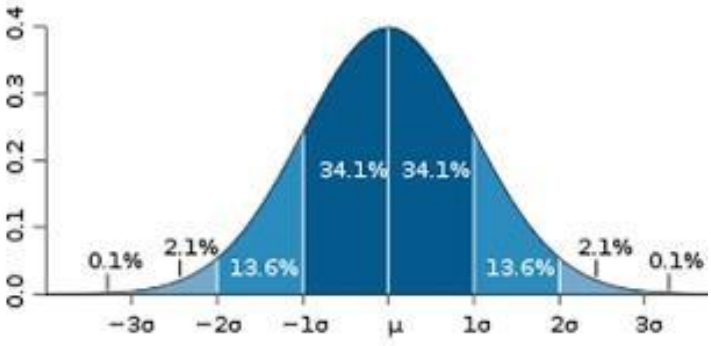
Przydatne pakiety: **tidyverse, zoo, caret**

Zadanie 1. Usuwanie błędów grubych

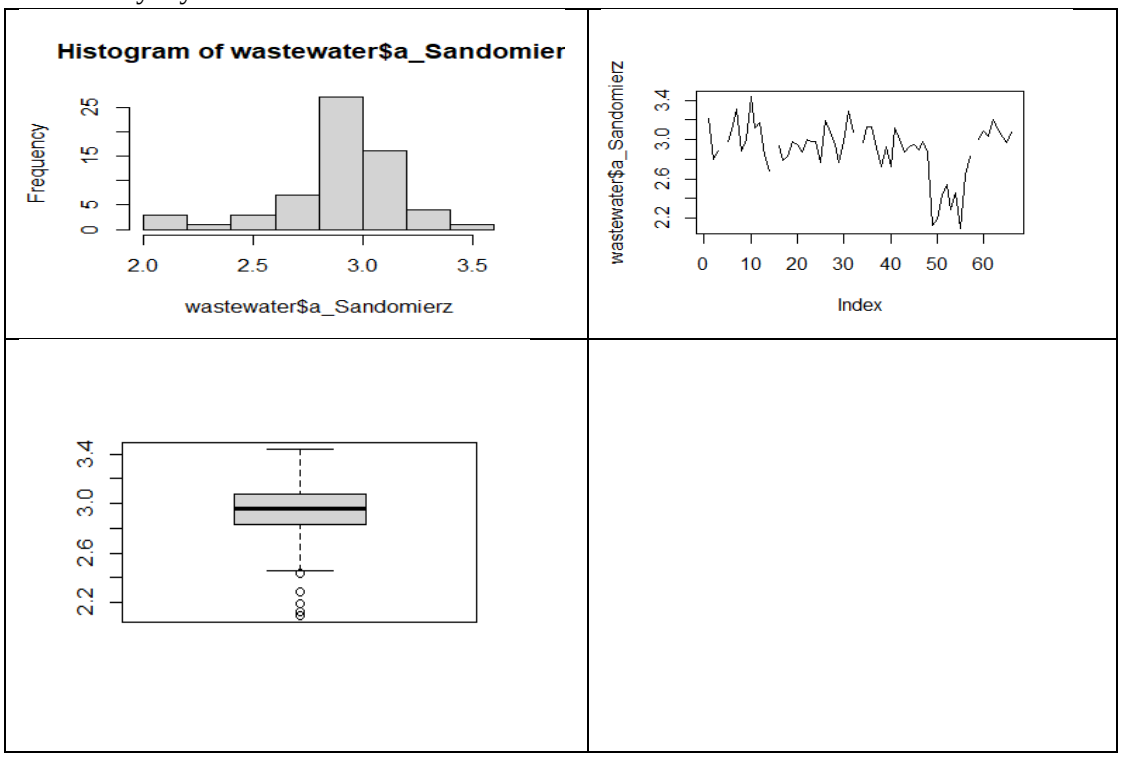
a) Wczytaj dane z pliku wastewater.txt do zmiennej wastewater.
b) Sprawdź strukturę danych, liczbę obserwacji, poprawność importu danych.
c) Wyszukaj informacje o danych.
d) Wykonaj wykresy sprawdzające występowanie wartości błędnych i ekstremalnych
   - boxplot(wastewater$Sandomierz)
   - plot(wastewater$Sandomierz, type="l")
   - hist(wastewater$Sandomierz, breaks=8)

   na podstawie wykresów oceń potrzebę wykonania preprocessingu- usuwania wartości sprzecznych z logiką i wartości ekstremalnych.

e) Maksymalny przepływ ścieków jaki może przyjąć oczyszczalnia to 20m^3/dobę. Za pomocą funkcji mutate utwórz mową zmienną a_Sandomierz, która będzie przechowywać wyłącznie wartości bez błędów grubych (wartości poniżej 0 i powyżej 20).
f) Zmodyfikuj zmienną a_Sandomierz, dla wartości, które możemy określić jako ekstremalne. Zastąp te wartości NA.



g) Ponownie wykonaj wizualizacji dla utworzonej zmiennej po preprocessingu.
Oczekiwany wynik:



Zadanie 2. Zastępowanie braków danych

a) Przypomnij sobie zasady zastępowania braków tabel w zależności od typu danych
   - Tabela, dane ilościowe,
   - Tabela, dane jakościowe,
   - Szereg czasowy, dane ilościowe, w zależności od składowych szeregu,
   - Szereg czasowy, dane ilościowe, w zależności od składowych szeregu.

b) Zastąp braki danych (NA), w danych z poprzedniego zadania wybraną przez siebie metodą. W przypadku szeregów czasowych z trendem lub sezonowością, lub obiema tymi składowymi przydatny jest pakiet „zoo" i np. funkcjami: na.approx(), na.fill(), na.StructTS(). Zapoznaj się z dokumentacją tych funkcji oraz „imputeTS" z funkcjami: na.mean(), na_replace(), na_random(), na_interpolation(), na_ma().
   Użycie niektórych funkcji dla szeregów czasowych wymaga zmiany df w szereg czasowy.
   Aby ocenić sezonowość i trend skorzystaj z wykresów ACF i PACF oraz wykresu liniowego.

c) Wykonaj raport w Markdown zawierający cele analizy, sposób zastępowania oraz interpretację.

Zadanie 3. Wykonaj przekształcenia danych

a) Wykonaj standaryzację danych i zapisz jako zmienną sand_std,
b) Wykonaj przekształcenie min-max i zapisz jako zmienną sand_min,
c) Wykonaj przekształcenie box-cox i zapisz jako zmienną sand_box,
d) Wykonaj logarytmowanie danych i zapisz jako zmienną sand_log,

e) Wykonaj różnicowanie danych (o 1) i zapisz jako zmienną sand_lag.

Zapisz przekształcenia do nowego df *sand_trans*.

Przydatne funkcje preProcess() z pakietu „caret". Zobacz jak wyglądają statystyki opisowe i histogramy dla przekształconych danych.

Zadanie 4. Wykonaj preprocessing danych tabelarycznych

Wczytaj dane *cars* i wykonaj preprocessing danych. Usuń wartości sprzeczne z logiką i wartości ekstremalne. Usuń duplikaty jeśli istnieją. Wykorzystaj funkcje z pakietów tidyverse, caret.

Zastąp braki danych stosując wykonywanie podzbiorów.

Przygotuj raport RMarkdown wraz z komentarzem wyboru metod oraz uzyskanych wyników.