Eksploracja danych
Lista nr 1
10 III 2025

Wprowadzenie
Naszym celem jest zastosowanie poznanych metod analizy opisowej (wykresy i wskaźniki sumaryczne) w analizie wybranych danych oraz szczegółowa interpretacja otrzymanych wyników. Wykorzystujemy zbiór danych `WA_Fn-UseC_-Telco-Customer-Churn.csv`, który zawiera informacje o klientach pewnej firmy telekomunikacyjnej (źródło: https://www.kaggle.com/datasets/blastchar/telco-customer-churn). Tego rodzaju dane są wykorzystywane w analizie migracji klientów (ang. attrition/churn analysis). W dużym uproszczeniu, firma zbiera dane o zachowaniu swoich klientów, próbując przewidzieć, kto może odejść i następnie stara się wprowadzić odpowiednie działania zachęcające do pozostania.

Opis danych
Podstawowy opis danych dostępny jest na portalu Kaggle.

Sprawozdanie
Z tej listy obowiązuje przygotowanie sprawozdania! Do sprawozdania (plik w formacie PDF) należy dołączyć plik źródłowy w formacie RNW lub RMD i ewentualnie kody źródłowe dodatkowych funkcji lub skryptów w R (jeżeli takie były wykorzystane). Uwaga: R-kody powinny być pisane starannie i zawierać komentarze! Pozostałe informacje/zalecenia dotyczące formy sprawozdania można znaleźć w szablonie dostępnym na stronie kursu na ePortalu PWr.

Prezentacja wyników
Przeprowadzając analizę, należy wziąć pod uwagę wszystkie zmienne/cechy zawarte w analizowanych danych, ale w sprawozdaniu proszę umieszczać tylko najważniejsze/najciekawsze wyniki, tzn. dla wybranych zmiennych jakościowych i ilościowych. Ważna jest także odpowiednia graficzna prezentacja otrzymanych rezultatów. W szczególności zachęcam do poszukania i wykorzystania ciekawych i mniej standardowych wykresów (patrz np. https://r-graph-gallery.com/).

Etap 1 Przygotowanie danych. Podstawowe informacje o danych.

a) Wczytaj zbiór danych `WA_Fn-UseC_-Telco-Customer-Churn.csv` do przestrzeni roboczej R. Po wczytaniu upewnij się czy typy poszczególnych zmiennych zostały prawidłowo rozpoznane (zmienne jakościowe – typ *factor*, zmienne ilościowe – typ *numeric*).

b) Przyjrzyj się danym i odpowiedz:

- Jaki jest rozmiar danych (liczba przypadków i cech)?
- Jakie są typy poszczególnych cech?
- Czy wszystkie cechy występujące w danych będą przydatne w analizie? W szczególności: czy w danych występuje cecha(y), które pełnią rolę identyfikatorów klientów? Jeżeli tak, to wskaż te cechy i usuń je przed dalszą analizą.
- Czy w danych występują brakujące obserwacje? Jeżeli tak, to jak są one kodowane?
- Czy w danych występują nietypowe wartości (np. niestandardowe kodowanie brakujących obserwacji)?

Wskazówka: do przestrzeni roboczej R dane można wczytać wykonując polecenie:

```
> dane <- read.csv(file="WA_Fn-UseC_-Telco-Customer-Churn.csv", stringsAsFactors = TRUE)
```

ETAP 2 Analiza opisowa — wskaźniki sumaryczne i wykresy

a) Wyznacz podstawowe wskaźniki sumaryczne (miary położenia i rozproszenia) dla poszczególnych cech/zmiennych.

b) Zilustruj rozkład poszczególnych cech/zmiennych za pomocą odpowiednich wykresów (wykresy słupkowe, histogramy, wykresy pudełkowe itp.).

c) Wykorzystując wykresy rozrzutu dla par zmiennych ciągłych, przeanalizuj czy występują istotne (w szczególności liniowe) zależności pomiędzy zmiennymi.

d) Zinterpretuj otrzymane rezultaty, w szczególności odpowiadając na pytania:

   - Jaki jest zakres możliwych wartości dla poszczególnych zmiennych?
   - W przypadku zmiennych ilościowych:
     i. Czy wszystkie zmienne mają rozkład symetryczny?
     ii. Które cechy charakteryzują się największą zmiennością?
   - W przypadku zmiennych jakościowych:
     i. Co można powiedzieć o częstości przyjmowania poszczególnych kategorii?

ETAP 3 Analiza opisowa z podziałem na grupy

a) Przeprowadź analizę opisową, podobnie jak w etapie 2 (wyznaczając wskaźniki opisowe i podstawowe wykresy), tym razem jednak z podziałem na dwie grupy:

   - **grupa 1:** `Churn=='Yes'` – grupa klientów, którzy odeszli (zrezygnowali),
   - **grupa 2:** `Churn=='No'` – grupa lojalnych klientów, którzy korzystają dalej z usług firmy.

b) Analizując wyniki z punktu a), odpowiedz na pytanie: które ze zmiennych wykazują największe zróżnicowanie wartości/rozkładu w grupach, tzn. pozwalają na najlepsze rozróżnienie klientów z dwóch grup?

ETAP 4 Podsumowanie – wnioski z przeprowadzonej analizy

a) Podsumuj krótko wyniki otrzymane w etapach 1-3. Jakie wnioski można wyciągnąć na podstawie przeprowadzonych analiz?

b) Czego dowiedziałeś(aś) się o klientach tej firmy, przeprowadzając analizę danych? Czym, przede wszystkim, charakteryzują się klienci? Z jakich usług głównie korzystają?

c) Na podstawie wyników etapu 3-ego spróbuj odpowiedzieć na pytania: Jaka jest główna przyczyna odchodzenia klientów (inaczej: co charakteryzuje klientów, którzy rezygnują z usług firmy?) Co (Twoim zdaniem) firma powinna zrobić, aby przeciwdziałać odchodzeniu klientów?