Projekt zaliczeniowy

Państwa zadaniem jest przeanalizowanie jednego ze zbiorów danych, dotyczących wynajmu lokali poprzez platformę Airbnb (http://insideairbnb.com/get-the-data.html). Na warsztat proszę wziąć zbiór, zawierający dane szczegółowe (*listings.csv.gz*) dla miasta, którego nazwa zaczyna się tą samą literą, co Państwa nazwisko (w przypadku braku takiego miasta, proszę wybrać dowolne z listy). Zakres analizy powinien obejmować co najmniej następujące kroki:

1. Poprawne załadowanie danych ze źródła internetowego do ramki danych, z uwzględnieniem nagłówków, kodowania zbioru, separatorów itd.;
2. Poznanie rozmiaru zbioru danych (liczby obserwacji i liczby zmiennych, które je opisują) i oszacowanie czasochłonności procesu analizy;
3. Wyświetlenie próbki surowych danych w celu wyrobienia sobie wyobrażenia o nich – poznania struktury danych i wstępnej oceny przydatności poszczególnych zmiennych;
4. Weryfikacja typów poszczególnych zmiennych (całkowite, zmiennoprzecinkowe, kategoryczne porządkowe, kategoryczne nominalne, zmienne typu logicznego, daty) i ich ewentualna korekta (zamiana typu *string* na *float*, interpretacja zmiennych numerycznych jako kategorii itp.);
5. Zbudowanie podsumowania zmiennych numerycznych opisujących zbiór, w postaci jednej tabelki, zawierającej podstawowe informacje, takie jak:
    a. wartości minimalne,
    b. wartości maksymalne,
    c. średnia,
    d. mediana,
    e. drugi (dolny) kwartyl,
    f. trzeci (górny) kwartyl,
    g. odchylenie standardowe,
    h. liczba danych brakujących lub nienumerycznych.

    W tym kroku należy również dokonać analogicznej analizy zmiennych kategorycznych, dającej dla każdej z nich informacje m.in. takie jak:

    a. liczby poszczególnych kategorii i ich liczności,
    b. wartości najczęściej występującej i częstości jej występowania,
    c. liczba wartości unikalnych,
    d. liczba braków danych.
6. Sprawdzenie, czy w zbiorze występują braki danych. Należy sporządzić odrębne podsumowanie, skupiając się na poszukiwaniu brakujących wartości w zbiorze – Pozwoli to Państwu odpowiedzieć na pytanie, jakie zmienne zawierają braki i jaka jest ich liczba, z czego mogą one wynikać itd.

    Etap ten (wraz z poprzednim) pozwoli Państwu odnaleźć błędy w danych – brakujące wartości, błędne interpretacje rodzaju zmiennych itp. Da również wskazówki, które atrybuty wybrać do analizy (pod kątem ich istotności dla przewidywań modelu), czy i jak uzupełnić brakujące dane (ewentualnie usunąć wiersze/kolumny, zawierające zbyt wiele braków danych), dokonać ich transformacji itd.

7. Wizualizacja rozkładu (wybranych) zmiennych (zarówno numerycznych, jak i kategorycznych) poprzez histogramy i próba ich scharakteryzowania (np. poprzez ich skośność i kurtozę) – będzie to pomocne np. w procesie imputacji (uzupełniania) zmiennych numerycznych;

8. Przeprowadzenie czyszczenia danych, obejmujące m.in.:

    a. uzupełnienie brakujących danych (np. wartością stałą, średnią/medianą/modą dla całego zbioru lub dla podzbiorów według kategorii, poprzez interpolację itp.), usunięcie wierszy/kolumn, zawierających zbyt wiele braków danych,
    b. przycięcie odstających wartości (ang. *outliers*) – pomocne będą m.in. takie techniki, jak wykres punktowy (gdzie nanosimy na obu osiach ten sam atrybut) lub wykres pudełkowy i ewentualna normalizacja danych numerycznych (metodą *min-max* lub *Z-score*) – niektóre algorytmy modelowania danych są wrażliwe na punkty odstające (np. metody regresji liniowej, korelacja Pearsona) czy różnice w zakresie zmienności poszczególnych atrybutów (niektóre algorytmy klasyfikacji bądź grupowania);

9. Zbadanie zależności pomiędzy zmiennymi – krok ten pozwoli odkryć związki pomiędzy poszczególnymi zmiennymi; informacje te mogą także zostać użyte, np. na etapie transformacji zmiennych lub do podjęcia decyzji, które zmienne wybrać do budowy modelu:

    a. obliczenie macierzy korelacji (można użyć współczynnika korelacji rang Spearmana lub współczynnika Pearsona) pomiędzy zmiennymi numerycznymi i zwizualizowanie ich za pomocą wykresów punktowych (ang. *scatter plots*) lub tzw. wykresów par zmiennych (ang. *pairplots*),

    b. ewentualne zbadanie zależności pomiędzy zmiennymi kategorycznymi (współczynnik *V* Cramméra) i zależności pomiędzy zmiennymi kategorycznymi i numerycznymi (współczynnik *R* modelu liniowego z jedną zmienną kategoryczną, która objaśnia zmienną numeryczną) oraz (podobnie jak powyżej) zwizualizowanie tych zależności w formie wykresów;

10. Opracowanie wyników analizy – bądź w postaci samodzielnego raportu w formie pliku PDF (ale zawierającego kod, użyty do jej sporządzenia), bądź – lepiej – w formie notatnika Jupyter Notebook / Google Colab i udostępnienie go (np. na GitHub-ie lub Google Drive).