SPRAWOZDANIE

Zajęcia: Nauka o danych I

Prowadzący: prof. dr hab. Vasyl Martsenyuk

|  |  |
| --- | --- |
| Laboratorium Nr 5  Data 23.11.2024  Temat: "Wykorzystanie narzędzi do eksploracyjnej  analizy danych (EDA)"  Wariant 10 | Anna Więzik  Informatyka  II stopień, niestacjonarne,  1 semestr, gr.1b |

1. Polecenie:

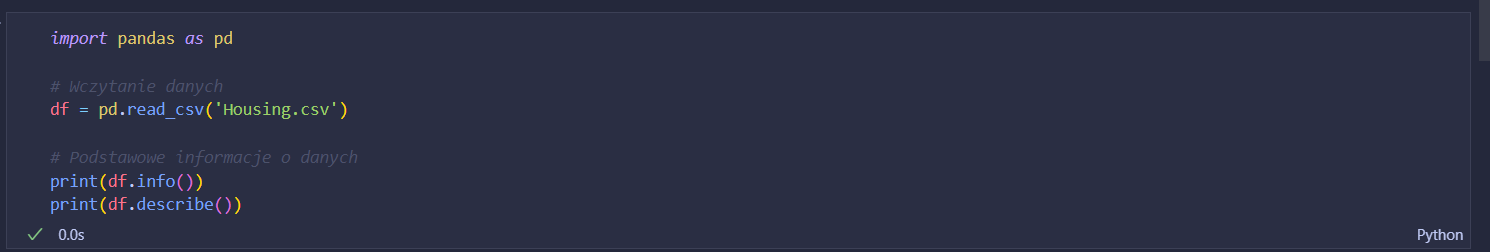
Premise Child Health COVID-19 Health Services Disruption Survey 2020

<http://ghdx.healthdata.org/record/ihme-data/premise-child-health-covid-19-health-services-disruption-survey-2020>

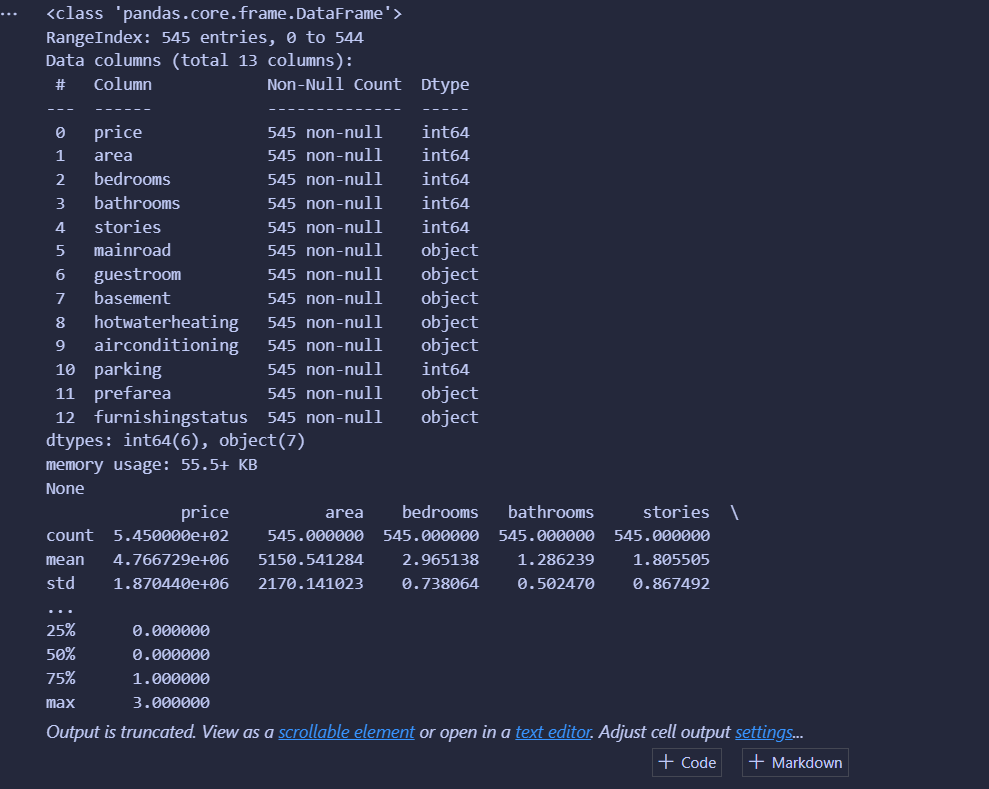
Link do repozytorium: <https://github.com/AnaShiro/NoD1_2024>

1. Opis programu opracowanego

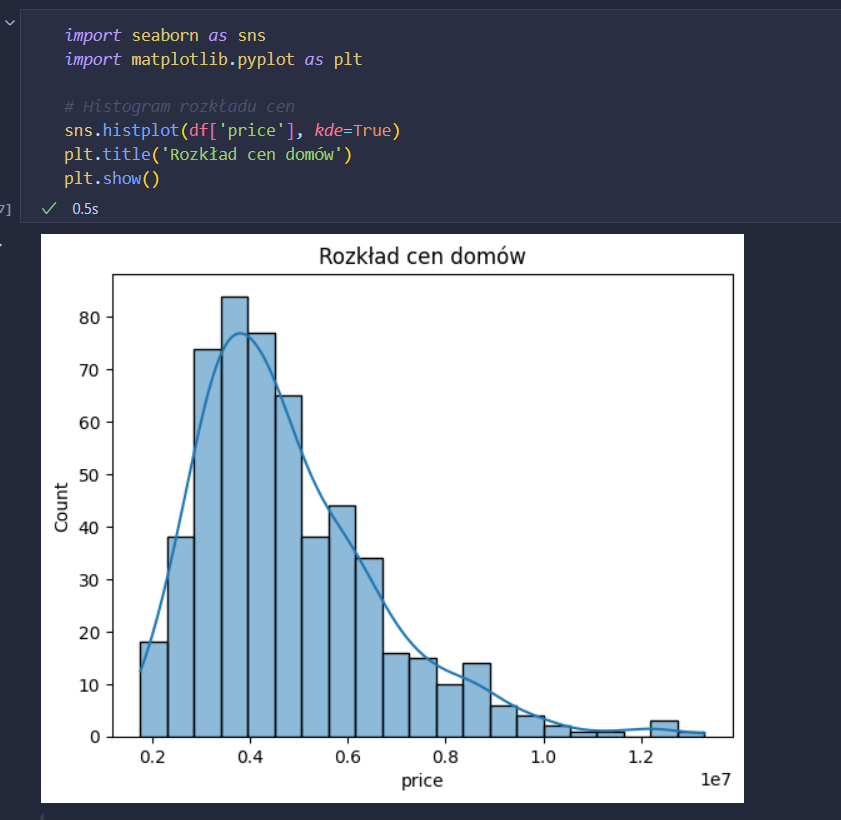
* Przygotowanie środowiska pracy

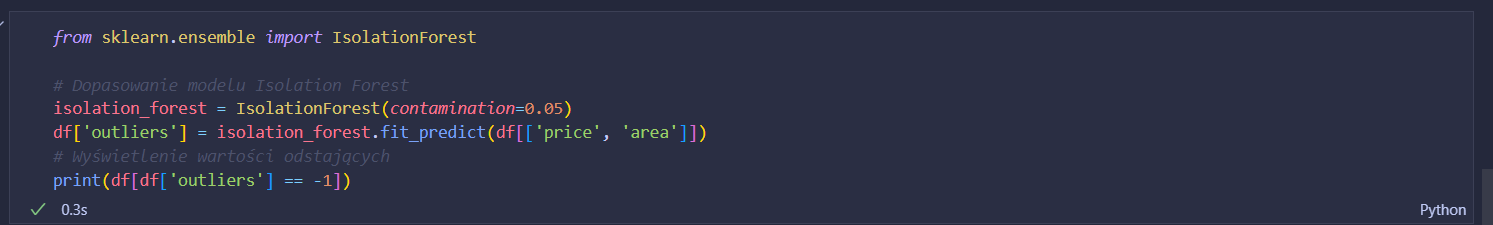


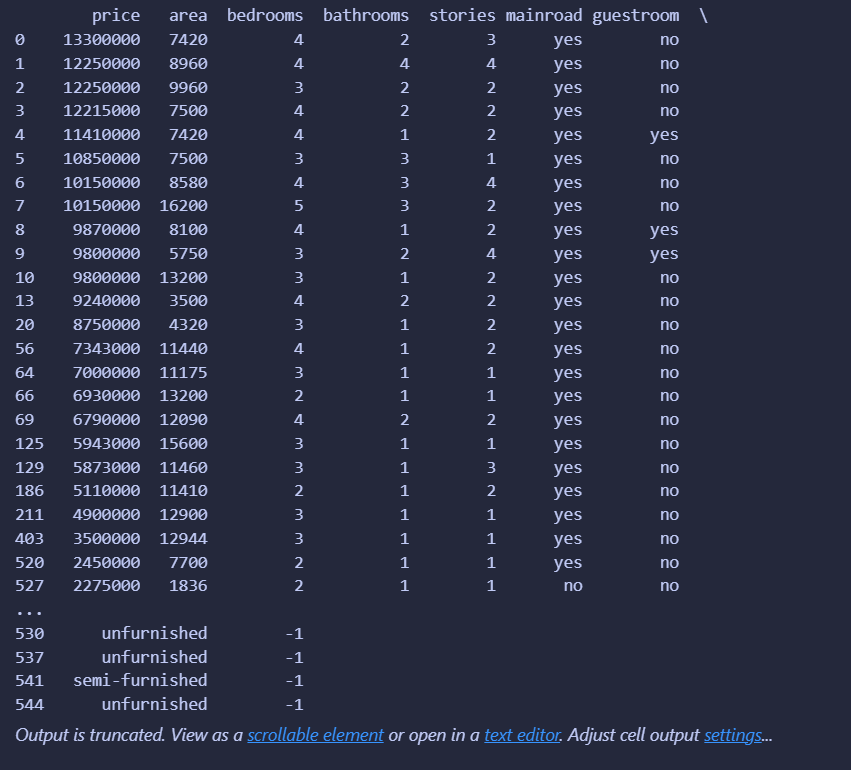
* Wczytanie i wstępne przetwarzanie danych



* Detekcja wartości odstających







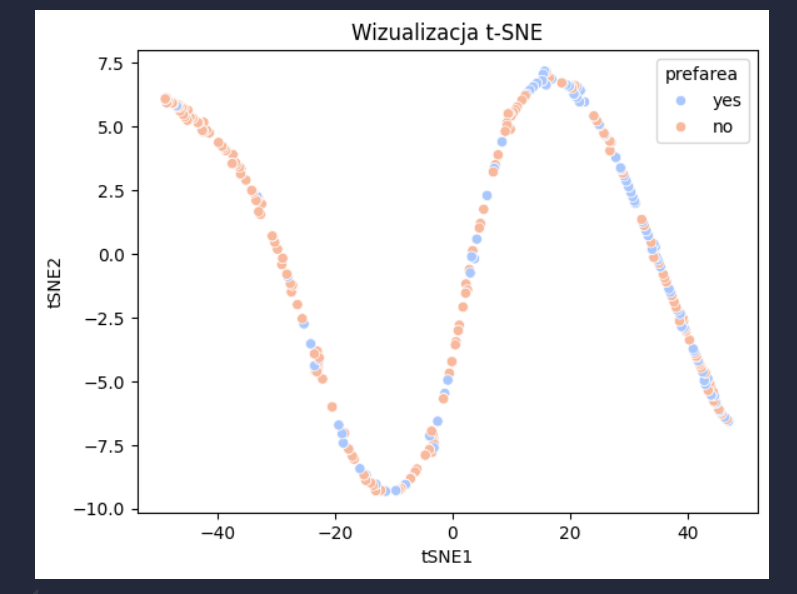


* Analiza głównych składowych (PCA)
* Wizualizacja redukcji wymiarowości- t-SNE

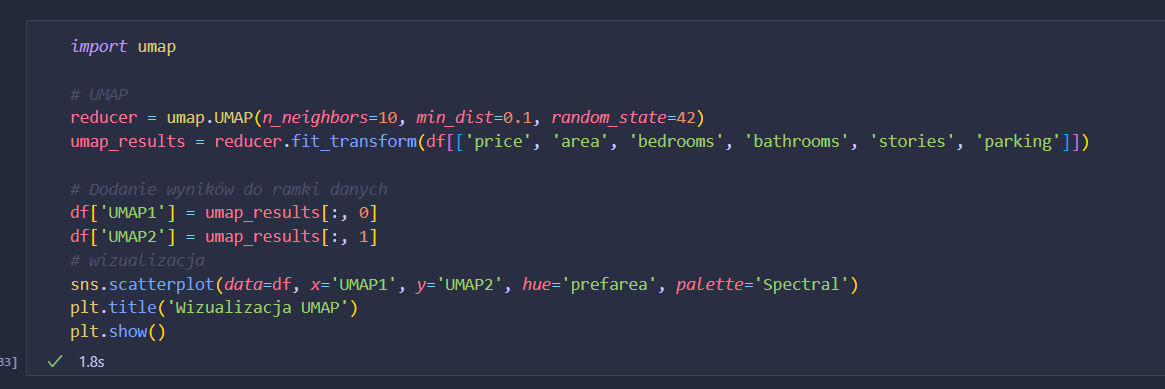


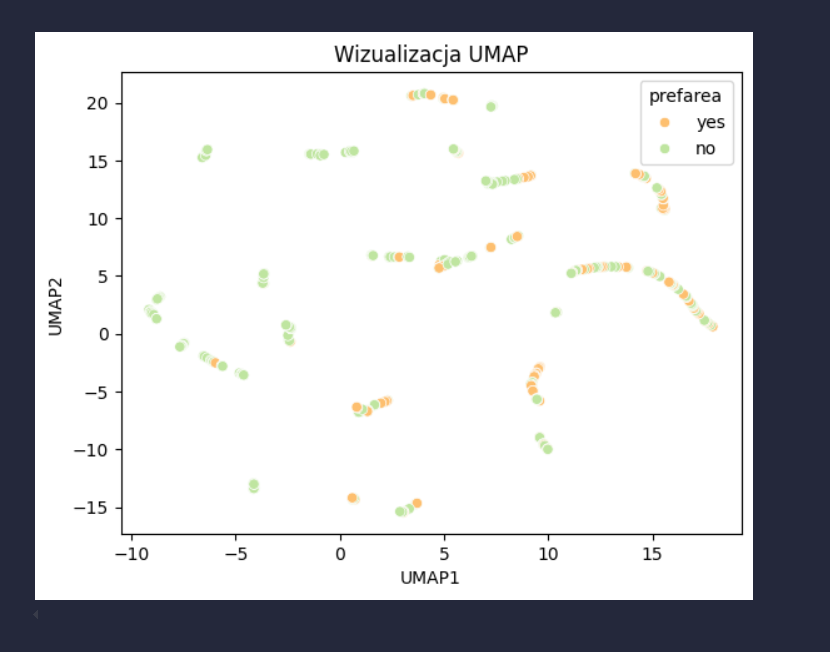






* Wizualizacja redukcji wymiarowości-UMAP





* Testy statystyczne





1. Pytania kontrolne

* Jak działa algorytm Isolation Forest i jak interpretować jego wyniki?

Algorytm Isolation Forest wykrywa anomalie przez "izolowanie" rzadkich punktów w danych, budując losowe drzewa decyzyjne. Wyniki interpretujemy na podstawie tzw. score anomalności — punkty z wysokim wynikiem (bliskim 1) są anomaliami, a te z wynikiem bliskim 0 to dane normalne.

* W jaki sposób analiza PCA może pomóc w eksploracyjnej analizie danych?

PCA (Principal Component Analysis) redukuje wymiarowość danych, przekształcając je w nowe, niezależne zmienne (główne składowe), co ułatwia ich wizualizację i wykrywanie ukrytych wzorców.

* Jakie są zalety wykorzystania interaktywnych wizualizacji?

Interaktywne wizualizacje umożliwiają użytkownikom dynamiczne eksplorowanie danych, pozwalając na łatwiejsze odkrywanie wzorców, zrozumienie zależności i szybsze podejmowanie decyzji.

* Jak interpretować wyniki testu ANOVA?

ANOVA porównuje średnie wartości w różnych grupach. Jeśli wynik testu (p-wartość) jest mniejszy niż 0,05, oznacza to, że przynajmniej jedna grupa różni się statystycznie od innych.

* Jak działa algorytm t-SNE i kiedy warto go stosować?

t-SNE (t-Distributed Stochastic Neighbor Embedding) jest metodą redukcji wymiarowości, która zachowuje lokalne struktury danych, idealna do wizualizacji skomplikowanych danych w niskich wymiarach (np. 2D, 3D).

* W jaki sposób algorytm UMAP różni się od t-SNE?

UMAP (Uniform Manifold Approximation and Projection) jest podobny do t-SNE, ale jest szybszy i skalowalny na większe zbiory danych. UMAP zachowuje zarówno lokalne, jak i globalne struktury danych.

* Jak interpretować macierz korelacji?

Macierz korelacji przedstawia zależności między zmiennymi. Wartości bliskie 1 lub -1 wskazują na silną pozytywną lub negatywną korelację, natomiast wartości bliskie 0 oznaczają brak zależności.

1. Wnioski

t-SNE jest potężnym narzędziem do redukcji wymiarowości i wizualizacji danych. Dzięki swojej zdolności do zachowania lokalnych relacji w danych o wysokiej wymiarowości, umożliwia odkrywanie klastrów i wzorców w sposób wizualnie intuicyjny. Jednak w interpretacji wyników należy uwzględniać ograniczenia dotyczące globalnych relacji i konieczności dostosowania hiper parametrów.

Wizualizacje UMAP są cennym narzędziem do interpretacji danych o wysokiej wymiarowości, dostarczając intuicyjnego, niskowymiarowego przedstawienia. Poprzez zachowanie zarówno lokalnych sąsiedztw, jak i globalnych relacji, UMAP pozwala wizualnie wykrywać istotne struktury i związki w złożonych zbiorach danych.

ANOVA jest potężnym narzędziem do analizy różnic między grupami, a funkcja anova.lm w Pythonie pozwala na łatwe przeprowadzenie tej analizy. Kluczową rolę w interpretacji wyników odgrywają wartości p oraz F, które wskazują, czy różnice pomiędzy grupami są istotne statystycznie.