Praca domowa

Paweł Strawiński

28 listopada 2023

Firma dla której pracujesz opracowuje model, za pomocą którego będzie analizowany wpływ czynników społecznych i ekonomicznych za zapadalność na nowotwory.

Twoim zadaniem jest przygotowanie raportu z procesu oczyszczenia zbioru danych z obserwacji odstających, które według Twojej najlepszej wiedzy mogą zaszkodzić budowanemu modelowi.

Dane są podzielone na dwa zbiory: zbiór `socioeconomic.csv` zawiera charakterystyki dotyczące zapadalności na choroby, oraz charakterystyki społeczno-ekonomiczne. Zbiór `geo_size.csv` zawiera informacje o przeciętnym rozmiarze gospodarstwa domowego oraz informacje o rejonie geograficznym (hrabstwo i stan USA). Kluczem łączenia obu zbiorów jest kolumna `index`.

Rozpocznij pracę od zapoznania się z danymi oraz ich uporządkowania. Zwróć uwagę, iż niektóre obserwacje mogą nie być kompletne.

Wskazówka: Oceniana będzie wartość merytoryczną opracowania. Liczba wykrytych obserwacji odstających nie ma znaczenia. Przede wszystkim zwracana będzie uwaga na logikę wywodu i argumentację uzasadniającą podjęte w trakcie analizy decyzje. Ważne decyzje powinny zostać zilustrowane odpowiednimi wykresami i/lub wartościami stosownych statystyk.

Nie ma błędnych odpowiedzi są mniej lub bardziej trafne.

Opracowanie powinno mieć formę raportu. Jesteś zobowiązana/zobowiązany dostarczyć raport w formie drukowanej (wydruk dwustronny) oraz elektronicznej (akceptowane formaty .pdf, ewentualnie .doc). Nieprzekraczalny termin dostarczenia raportu to 25 stycznia 2022 godzina 18.00. Jeśli go nie dotrzymasz praca domowa zostanie uznana niewykonaną. Prace należy przesłać na adres elektroniczny `pstrawinski@wne.uw.edu.pl`, a wersję drukowaną pozostawić na portierni budynku 00-241 Warszawa, Długa 44/50 wejście od ulicy Długiej. Opracowanie należy opatrzyć imieniem, nazwiskiem i numerem indeksu autora. Opracowania anonimowe nie będą brane pod uwagę[1].

Limit długości tekstu: 27000 znaków tekstu (15 stron), bez rysunków i tabel.

Informacje dodatkowe:

Źródło danych: data.world, licencja CC BY 4.0 Deed

Zbiór `geo_size.csv`

- index - identyfikator pozwalający na łączenie obserwacji
- statefips - 2 cyfrowy kod identyfikujący stan,
- countyfips - 3 cyfrowy kod identyfikujący hrabstwo,
- avg household size,
- geography - nazwa regionu geofraficznego

Zbiór `socioeconomic.csv`

- index - identyfikator pozwalający na łączenie obserwacji
- avganncount - przęciętna roczna liczba zachorowań na mowotwór,

[1]Praca zostanie uznana niewykonaną

- avgdeathsperyear - przeciętna roczna śmiertelność,
- target_deathrate - docelowa wartość współczynnika zgonów,
- incidencerate - częstotliwość występowania,
- medincome - mediana dochodu,
- popest2015 - oszacowanie liczby ludności w 2015 roku,
- povertypercent - procent żyjących w ubóstwie,
- studypercap - wydatki na badania zdrowotne na mieszkańca,
- binnedinc - przedział przeciętnego dochodu,
- medianage - mediana wieku,
- medianagemale - mediana wieku mężczyzn,
- medianagefemale - mediana wieku kobiet,
- percentmarried - procent osób żyjących w związkach,
- pctnohs18_24 - procent osób bez ukończonej szkoły średniej,
- pcths18_24 - procent osób z ukończoną szkołą średnią,
- pctsomecol18_24 - procent osób ze studiami w koledżu,
- pctbachdeg18_24 - procent osób z dyplomem licencjata (bachelor),
- pcths25_over - procent osób z dyplomem wśród osób w wieku 25 i więcej,
- pctbachdeg25_over - procent osób z dyplomem conajmniej licencjata w wieku 25 la i więcej,
- pctemployed16_over - procent zatrudnionych wśród osób w wieku 16 lat i więcej,

- pctunemployed16_over - procent bezrobotnych wśród osób w wielu 16 lat i więcej,
- pctprivatecoverage - procent osób z prywatnym ubezpieczeniem zdrowotnym,
- pctprivatecoverage_alone - procent osób z wyłącznie prywatnym ubezpieczeniem zdrowotnym,
- pctempprivcoverage - procent osób z czasowym prywatnym ubezpieczeniem zdrowotnym,
- pctpubliccoverage - procent osób z publicznym ubezpieczeniem zdrowotnym,
- pctpubliccoverage_alone - procent osób z wyłącznie publicznym ubezpieczeniem zdrowotnym,
- pctwhite - procent ludności białej (pochodzenia europejskiego),
- pctblack - procent ludności pochodzenia afrykańskiego/afroamerykańskiego,
- pctasian - procent ludności pochodzenia azjatyckiego,
- pctotherrace - procent ludności o innym pochodzeniu,
- pctmarriedhouholds - procent gospodarstw domowych z małżeństwami
- birthrate - stopa urodzeń