Zajęcie 1. Ustalenia platformu Jupyter. Użycie biblioteki pandas w celu eksploracji i wizualizacji danych

Abstract

Celem jest nabycie podstawowej znajomości języka Python rozwiązując zadanie tworzenia i wyświetlenia ramki danych odpowiednio do określonego wariantu

1. Instalacje ekosystemu

Możliwe są dwa sposoby:

1. ekosystem Anaconda (Python+7500 packages+utilities)
   Link do instalacji -
   `https://www.anaconda.com/products/individual`
   `https://docs.anaconda.com/anaconda/install/windows/`

2. "ręcznie":
   - Python 3 (pip - for installation packages)
   - Jupyter Lab (Jupyter Notebook)
   - miniconda (conda - for installation additional Anaconda packages)

2. Podstawowe umiejętności użycia języka Python

W celu zdobycia podstawowyh umiejętności Python nalerzy przećwiczyć tutoriały do zajęcia 1-14 umieszczone na e-uczelni.

3. Podstawowe umiejętności użycia biblioteki pandas

W celu zdobycia podstawowyh umiejętności biblioteki pandas nalerzy przećwiczyć tutoriały do zajęcia 15-16 umieszczone na e-uczelni.

4. Opis danych

Dane do zadania będą pobrane ze strony `http://ghdx.healthdata.org/ihme_data` która przedstawia sobą repozytorium danych socioekonomicznych.

One zawierają dane badań statystycznych w zakresie gospodarki, demogradii oraz służby zdrowia.

Dane mogą być przedstawione w postaci plików formatów .csv lub .xslx

5. Warianty Zadania

Zadanie dotyczy pobrania danych z pliku, tworzenia ramki danych, wykonania poszczególnych zadań poniżej na podstawie odpowiedniego zbioru danych:

Warianty

1. Development Assistance for Health Database 1990-2020 `http://ghdx.healthdata.org/record/ihme-data/development-assistance-health-database-1990-2`

2. Gross Domestic Product Per Capita 1960-2050 `http://ghdx.healthdata.org/record/ihme-data/global-gdp-per-capita-1960-2050`

3. Global Health Spending 1995-2018 `http://ghdx.healthdata.org/record/ihme-data/global-health-spending-1995-2018`

4. United States Healthcare Spending by Race and Ethnicity 2002-2016 `http://ghdx.healthdata.org/record/ihme-data/united-states-healthcare-spending`

5. Global Young People Smoking Prevalence and Initiation Age 1990-2019 `http://ghdx.healthdata.org/record/ihme-data/global-young-people-smoking-preva`

6. Global Burden of Disease Study 2019 (GBD 2019) Chewing Tobacco Use Prevalence 1990-2019 `http://ghdx.healthdata.org/record/ihme-data/gbd-2019-chewing-tobacco-use-prevalence-1990-2019`

7. Global Burden of Disease Study 2019 (GBD 2019) Smoking Tobacco Use Prevalence 1990-2019 `http://ghdx.healthdata.org/record/ihme-data/gbd-2019-smoking-tobacco-use-prevalence-1990-2019`

8. ORB General Population COVID-19 Health Services Disruption Survey 2020 http://ghdx.healthdata.org/record/ihme-data/orb-general-population-covid-19-

9. Premise Women's Health COVID-19 Health Services Disruption Survey 2020 http://ghdx.healthdata.org/record/ihme-data/premise-women%E2%80%99s-health-covid-19-health-services-disruption-survey-2020

10. Premise Child Health COVID-19 Health Services Disruption Survey 2020 http://ghdx.healthdata.org/record/ihme-data/premise-child-health-covid-1

11. Premise General Population COVID-19 Health Services Disruption Survey 2020 http://ghdx.healthdata.org/record/ihme-data/premise-general-populatio

12. United States COVID-19 Scenarios 2020-2021 http://ghdx.healthdata.org/record/ihme-data/united-states-covid-19-scenarios-2020-2021

13. Global Burden of Disease Study 2019 (GBD 2019) Relative Risks http://ghdx.healthdata.org/record/ihme-data/gbd-2019-relative-risks

14. United States Health-Care Spending Attributable to Modifiable Risk Factors 2016 http://ghdx.healthdata.org/record/ihme-data/united-states-health-c

15. Global Child Growth Failure Geospatial Estimates 2000-2019 http://ghdx.healthdata.org/record/ihme-data/global-child-growth-failure-geospatial-e

16. Low- and Middle-Income Country Drinking Water and Sanitation Facilities Access Geospatial Estimates 2000-2017 http://ghdx.healthdata.org/record/ihme-data/lmic-wash-access-geospatial-estimates-2000-2017

17. Low- and Middle-Income Country Oral Rehydration Therapy Coverage Geospatial Estimates 2000-2017 http://ghdx.healthdata.org/record/ihme-data/lmic-oral-rehydration-therapy-coverage-geospatial-estimates

18. Global Fertility, Mortality, Migration, and Population Forecasts 2017-2100 http://ghdx.healthdata.org/record/ihme-data/global-population-forecasts-

Zadania:

- ładowanie biblioteki Pandas
- tworzenie ramki danych ze słownika
- zachowanie ramki danych pobranych z pliku w formacie csv (xlsx)
- tworzenie ramki danych z listy list
- transponowanie (wymieniamy kolumny a wierszy)
- wyświetlić pierwsze 10 wierszy ramki danych
- wyświetlić ostatnie 10 wierszy ramki danych
- wyświetlić informację o ramce danych
- wyświetlić, ile wierszy i kolumn znajduje się w ramce danych
- wyświetlić informację statystyczną o kolumnach liczbowych (wartości niepowtarzalne, średnia, odchylenie standardowe, minimum, kwartyle, maksimum)
- wyświetlić informację statystyczną o kolumnach kategoryzowanych (ile unikalnych wartości, top - jaka jest najpopularniejsza wartość, freq - jak często najpopularniejsza)
- usunąć brakujące wartości w ramce danych

- przedstawić wybór wierszy i kolumny używając nazw oraz indeksów na różne sposoby

- przedstawić wybór wierszy z ramki danych pod warunkiem odnośnie określonej wartości kolumny

- przedstawić wybór wierszy z ramki danych pod warunkiem spełnienia kilku warunków jednocześnie

- wybrać wiersze które zawierają w kolumnie kategoryzowanej określone słowo

- wybrać wiersze które nie zawierają w kolumnie kategoryzowanej określone słowo

- utwórz kolumnę na podstawie istniejącyn

- usuń kolumnę

- zmień nazwę kolumny

- zachowaj ramkę danych jako plik csv na komputerze

- wyświetlić średnia (maksymalną, minimalną) wartość z jednej kolumny

- wyświetlić liczbę wierszy

- wyświetlić wartości unikatowe w kolumnie

- wyświetlić liczby rekordów odpowiadających do wartości
- sortowanie wierszy ramki danych według wartości określonej kolumny (malejąco, rosnąco)
- wyświetlić wierszy dla 10 największych (najmniejszych) wartości określonej kolumny
- wyświetlić wierszy dla 10 największych wartości określonej kolumny pod warunkiem określonych wartości innej kolumny
- grupowanie wierszy wedlug wartości kolumny kategoryzowanej, potem - uśrednienie wartości wszystkich kolumn w grupie - MultiIndex
- grupowanie wierszy wedlug wartości kolumny kategoryzowanej, potem - uśrednienie wartości dla pewnych kolumn, liczba wartości i mediana dla pozostałych kolumn w grupach
- wyświetlić nazwy kolumn indeksu złożonego
- sortować kolumnę indeksu złożonego
- stworzyć tabelę przystawną (pivot table) na podstawie ramki danych
- wyświetlić indeksy i kolumny tabeli przystawnej
- utwórz indeks złożony tabeli przystawnej i wyświetl go
- zaimportuj moduł pyplot z biblioteki matplotlib

- wskazać, że wykresy należy rysować bezpośrednio w zeszycie, a nie w osobnej zakładce
- wyświetlić wykres na podstawie tabeli przystawnej
- narysować histogram na podstawie wartości kolumny
- przedstawić sposoby łączenia ramek danych za pomocą metod merge i concat
- pokazać dodawanie nowych kolumn za pomocą operacji matematycznych
- przedstawić na przykładzie dodawanie nowych kolumn z pomocą funkcji lambda
- przedstawić możliwości pracy z dużymi plikami przy użyciu argumentu chunksize

Sprawozdania w postaci:

1. Sprawozdanie (plik .pdf)
2. plik .ipynb
3. pdf-eksport pliku .pynb

zachować w zdalnym repozytorium (np Github) link na który umieściś w sprawozdaniu. Sprawozdanie należy wysłać na e-uczelnię zgodnie z ustalonym terminem.

References

References

[pandasUG] Pandas User's Guide `https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html`

[DA2016] Data Analysis with Python and pandas using Jupyter Notebook https://dev.socrata.com/blog/2016/02/01/pandas-and-jupyter-notebook.html