Priprema podataka

Bruno Polonijo, asistent

Priprema podataka 1.

- Da bi počeli raditi s podacima prvo moramo doći do podataka.
- To možemo kroz specijalizirane stranice kao: https://www.kaggle.com/
- Nakon toga moramo učitati skinuti CSV

Priprema podataka 2.

- Uvozimo pandas za rad nakon toga uvozimo csv koji smo ranije skinuli

```
import pandas as pd #uvoz programskih biblioteka
df = pd.read_csv('C://Users/Abhothoh/Documents/WLD_RTFP_country_2023-10-02.csv') #učitava csv datoteku i sprema ju u varijablu df
```

Priprema podataka 3.

- Nakon toga si prikažemo podatke da vidimo kako izgledaju
- Postoje dva načina:

```
print(df) #služi prikazu podataka iz varijable
```

```
      Open  High   Low  Close  Inflation      country ISO3        date
0     0.53  0.54  0.53   0.53        NaN  Afghanistan  AFG  2007-01-01
1     0.53  0.54  0.53   0.53        NaN  Afghanistan  AFG  2007-02-01
2     0.54  0.54  0.53   0.53        NaN  Afghanistan  AFG  2007-03-01
3     0.53  0.55  0.53   0.55        NaN  Afghanistan  AFG  2007-04-01
4     0.56  0.57  0.56   0.57        NaN  Afghanistan  AFG  2007-05-01
...    ...   ...   ...    ...        ...          ...  ...         ...
4793  2.74  2.78  2.70   2.75      -0.28  Yemen, Rep.  YEM  2023-06-01
4794  2.79  2.83  2.75   2.81      -1.85  Yemen, Rep.  YEM  2023-07-01
4795  2.85  2.89  2.81   2.83      -3.17  Yemen, Rep.  YEM  2023-08-01
4796  2.86  2.97  2.82   2.97       1.68  Yemen, Rep.  YEM  2023-09-01
4797  3.06  3.11  2.98   2.98       3.76  Yemen, Rep.  YEM  2023-10-01

[4798 rows x 8 columns]
```

```
df.head() #drugi "atraktivniji" način prikaza podataka
```

	Open	High	Low	Close	Inflation	country	ISO3	date
0	0.53	0.54	0.53	0.53	NaN	Afghanistan	AFG	2007-01-01
1	0.53	0.54	0.53	0.53	NaN	Afghanistan	AFG	2007-02-01
2	0.54	0.54	0.53	0.53	NaN	Afghanistan	AFG	2007-03-01
3	0.53	0.55	0.53	0.55	NaN	Afghanistan	AFG	2007-04-01
4	0.56	0.57	0.56	0.57	NaN	Afghanistan	AFG	2007-05-01

Priprema podataka 4.

- Ukoliko u datasetu postoje prazni redovi poželjno ih je uklonit. To radimo s naredbom dropna

```
df_cleaned = df.dropna(how='any') # Uklanja sve redove iz DataFrame-a 'df' koji sadrže barem jednu NaN (nedostajuću) vrijednost
```

```
df_cleaned = df.dropna(how='any', subset=df.columns[:-3]) #Uklanja redove iz DataFrame-a 'df' koji imaju barem jednu NaN vrijednost u svim stupcima
#osim zadnja tri
```

```
df_cleaned = df.dropna(how='all', subset=df.columns[:-3])  # Ovo će ukloniti redove gdje su svi podaci (osim zadnja tri stupca) NaN.
```

Priprema podataka 5.

- Nakon toga možete promjeniti imena stupaca na neka koja vam više odgovaraju

```
df.columns = ['Open', 'High', 'Low', 'Close', 'Inflation', 'Country', 'Code', 'Date'] #Može se promjenit ime stupca
```

Priprema podataka 6.

- Ukoliko vam je potrebno možete dodat i novi stupac.
- U primjeru u kreirani stupac se zapisuje da li postoji stopa inflacije ukoliko postoji vrijendost veća od 0 u stupcu inflation

```
df['Positive_Inflation'] = df['Inflation'].apply(lambda x: 'da' if x > 0 else 'ne')
# Dodajte novi stupac 'Positive_Inflation' koji će imati vrijednosti 'da' ili 'ne'
# ovisno o tome je li inflacija pozitivna ili ne
```

Priprema podataka 7.

- Ukoliko želite možete promjenit tip podataka nekog stupca. Primjerice za datum:

```
# Pretvorite niz s datumima u DateTime objekt
df['Date'] = pd.to_datetime(df['Date'], format='%Y-%m-%d')
```

```
# Konvertirajte datume u format 'DD/MM/YYYY'
df['Date'] = df['Date'].dt.strftime('%d/%m/%Y')
```

Priprema podataka 8.

- Ukoliko želite stupac možete uklonit s naredbom:

```
# Pretpostavimo da želite izbrisati stupac pod nazivom 'Code'
df.drop(columns=['Code'], inplace=True)

# Ako želite izbrisati više stupaca odjednom, možete navesti imena stupaca u listi.
# Na primjer, da izbrišete stupce 'Code' i 'Open':
df.drop(columns=['Code', 'Open'], inplace=True)
```

Priprema podataka 9.

- Konačno spremite sve napravljene promjene

```
df.to_csv('C:/Users/Abhothoh/Documents/cleaned_data2.csv', index=False) #Nakon što ste obradili podatke, možete ih spremiti u novu CSV datoteku.
```

- Ili ako želite samo pročišćenu datoteku:

```
df_cleaned.to_csv('C:/Users/Abhothoh/Documents/cleaned_data2.csv', index=False) #Nakon što ste obradili podatke, možete ih spremiti u novu CSV datoteku.
```

Pitanja?