14. час: Табеларно представљени подаци у *CSV* датотекама

На овом часу ћемо говорити о:

1. учитавању података из табела које су припремљене у формату *CSV*; и
2. припреми података пре обраде.

14.1. Учитавање података из локалних *CSV* датотека

Видели смо у претходним примерима да се најмукотрпнији посао обраде података састоји у томе да се подаци унесу у табелу. То је досадан посао који се често састоји у томе да се подаци просто прекуцају. Табеле са којима смо се сретали су зато биле веома мале. Модерна обрада података се, међутим, све више усмерава на анализу *огромних* количина података (енгл. *big data*) и ту прекуцавање података не долази у обзир.

Подаци се данас углавном прикупљају аутоматски, и програми за прикупљање података генеришу велике табеле података које после треба обрађивати. Постоје разни формати за табеларно представљање података, а најједноставнији од њих се зове *CSV*, (од енгл. *comma separated values* што значи "вредности раздвојене зарезима").

CSV датотека је текстуална датотека у којој редови одговарају редовима табеле, а подаци унутар истог реда су раздвојени зарезима. На пример, у фолдеру *podaci* се налази датотека *StanovnistvoSrbije2017.csv* која изгледа овако:

```
Старост,Мушко,Женско
0,33145,31444
1,33252,31105
2,33807,31475
3,34076,31952
4,33436,31643
5,34278,32505
6,33773,31523
7,33892,32185
8,34706,32396
9,34519,32177
10,34017,32064
11,34947,33251
... (итд) ...
84,11450,18529
85 и више,44817,78323
```

Ова табела садржи процену броја становника Републике Србије по годинама на дан 31.12.2017. Први ред табеле представља заглавље табеле које нам каже да табела има три колоне (Старост, Мушко, Женско). Врста

```
7,33892,32185
```

значи да се процењује да је 31.12.2017. у Србији било 33.892 седмогодишњих дечака и 32.185 седмогодишњих девојчица.

Библиотека `pandas` има функцију `read_csv` која учитава *CSV* датотеку и од ње прави табелу типа *DataFrame*. Ево примера:

In [1]:
```
# изврши ову ћелију
import pandas as pd
stanovnistvo = pd.read_csv("podaci/StanovnistvoSrbije2017.csv")
```

Пошто је табела велика, приказаћемо само првих неколико редова. Функција `head(N)` приказује првих N редова табеле (енгл. *head* значи "глава"):

In [11]:
```
# изврши ову ћелију
stanovnistvo.head(5)
```

Out[11]:

	Старост	Мушко	Женско
0	0	33145	31444
1	1	33252	31105
2	2	33807	31475
3	3	34076	31952
4	4	33436	31643

Функција `tail(N)` приказује последњих N редова табеле (енгл. *tail* значи "реп"):

In [12]:
```
# изврши ову ћелију
stanovnistvo.tail(5)
```

Out[12]:

	Старост	Мушко	Женско
81	81	16552	25345
82	82	15025	23036
83	83	13522	21435
84	84	11450	18529
85	85 и више	44817	78323

Табелу ћемо индексирати колоном "Старост":

In [14]:
```
# изврши ову ћелију
stanovnistvo1 = stanovnistvo.set_index("Старост")
stanovnistvo1.head(5)
```

Out[14]:

Старост	Мушко	Женско
0	33145	31444
1	33252	31105
2	33807	31475
3	34076	31952
4	33436	31643

In [15]:
```
# изврши ову ћелију
stanovnistvo1.tail(5)
```

Out[15]:

Старост	Мушко	Женско
81	16552	25345
82	15025	23036
83	13522	21435
84	11450	18529
85 и више	44817	78323

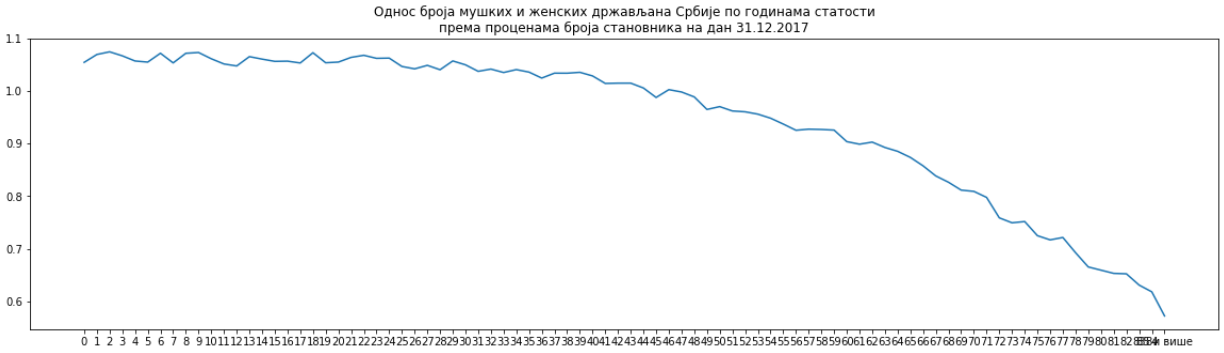
Сада ћемо урадити малу демографску анализу: израчунаћемо однос броја мушкараца и жена по годинама старости и приказаћемо податке хистограмом.

Прво ћемо табели додати нову колону "М/Ж" и у ту колону уписати израчунате односе:

In [19]:
```
# изврши ову ћелију
stanovnistvo1["М/Ж"] = 0.0
for i in stanovnistvo1.index:
    stanovnistvo1.loc[i, "М/Ж"] = stanovnistvo1.loc[i, "Мушко"] / stanovnistvo1.l
```

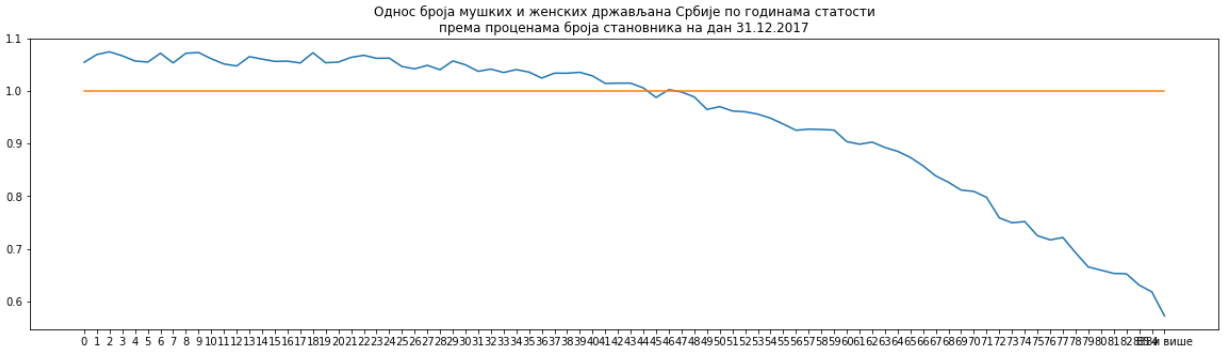
Потом ћемо приказати дијаграм:

In [23]:
```
# изврши ову ћелију
import matplotlib.pyplot as plt
plt.figure(figsize=(20,5))
plt.plot(stanovnistvo1.index, stanovnistvo1["М/Ж"])
plt.title("Однос броја мушких и женских држављана Србије по годинама статости\npp
plt.show()
plt.close()
```



Додаћемо линију на висини 1.0 да бисмо лакше уочили у ком тренутку број мушкараца постаје мањи од броја жена:

In [32]:
```
# изврши ову ћелију
plt.figure(figsize=(20,5))
plt.plot(stanovnistvo1.index, stanovnistvo1["М/Ж"])
plt.plot(stanovnistvo1.index, [1.0] * len(stanovnistvo1.index))
plt.title("Однос броја мушких и женских држављана Србије по годинама статости\npp
plt.show()
plt.close()
```



14.2. Трансформација табела пре обраде података

Често се дешава да табела са подацима нема заглавље. Тада се прилиом учитавања то мора нагласити функцији `read_csv` тако што се наведе `header = None` .

На пример, у фолдеру *podaci* се налази датотека *TemperaturneAnomalije.csv* која садржи податке о томе за колико степени Целзијуса је средња измерена температура на Земљи већа од оптималне у последњих 40 година. Ова табела има два дугачка реда који изгледају овако:

```
1977,1978,1979,1980,1981,...
0.22,0.14,0.15,0.3,0.37,...
```

У првом реду се налазе годину (1977--2017), а у другом измерена температурна аномалија. Видимо да табела нема заглавље. Зато ћемо је учитати на следећи начин:

In [6]:
```
# изврши ову ћелију
temp_anomalije = pd.read_csv("podaci/TemperaturneAnomalije.csv", header = None)
temp_anomalije
```

Out[6]:

	0	1	2	3	4	5	6	7	8	9	...	
0	1977.00	1978.00	1979.00	1980.0	1981.00	1982.00	1983.0	1984.00	1985.00	1986.00	...	2008
1	0.22	0.14	0.15	0.3	0.37	0.15	0.4	0.23	0.14	0.28	...	0

2 rows × 41 columns

Да бисмо добили податке у облику који се лакше обрађује транспоноваћемо табелу и колонама тако транспоноване табеле дати одговарајућа имена.

In [8]:
```
# изврши ову ћелију
temp_anomalije1 = temp_anomalije.T
temp_anomalije1.columns = ["Година", "Аномалија"]
```

Ево првих неколико редова табеле:

In [9]:
```
# изврши ову ћелију
temp_anomalije1.head(10)
```

Out[9]:

	Година	Аномалија
0	1977.0	0.22
1	1978.0	0.14
2	1979.0	0.15
3	1980.0	0.30
4	1981.0	0.37
5	1982.0	0.15
6	1983.0	0.40
7	1984.0	0.23
8	1985.0	0.14
9	1986.0	0.28

Табелу ћемо индексирати колоном "Година":

In [10]:
```
# изврши ову ћелију
temp_anomalije2 = temp_anomalije1.set_index("Година")
temp_anomalije2.head(5)
```

Out[10]:

Година	Аномалија
1977.0	0.22
1978.0	0.14
1979.0	0.15
1980.0	0.30
1981.0	0.37

Приказаћемо температурне аномалије дијаграмом:

In [13]:
```
# изврши ову ћелију
import matplotlib.pyplot as plt
plt.figure(figsize=(15,5))
plt.plot(temp_anomalije2.index, temp_anomalije2["Аномалија"], color="r")
plt.title("Температурне аномалије у периоду 1977--2017")
plt.show()
plt.close()
```

14.3. Задаци

Задатак 1. У фолдеру *podaci* се налази датотека *StanovnistvoSrbije2017.csv* (која има заглавље). Табела има три колоне које се зову "Старост", "Мушко" и "Женско".

(а) Учитати датотеку у структуру података *DataFrame* и индексирати табелу колоном "Старост".

(б) Додати табели нову колону "УкупноСт" и онда израчунати и у ту колону уписати податак о томе колики је укупан процењени број становника по старости. Приказати укупан процењени број становника по старости линијским дијаграмом.

(в) Додати табели нову врсту "УкупноПол" и онда израчунати и у ту врсту уписати податак о томе колики је укупан процењени број становника по полу. Приказати укупан процењени број становника по полу секторским дијаграмом.

Задатак 2. Ученици једног разреда су скакали у даљ. Сваки ученик је скако три пута и резултати су дати у датотеци *SkokUDalj.csv* која се налази у фолдеру *podaci*. Табела има заглавље и састоји се од четири колоне: "Презиме и име", "Скок1", "Скок2" и "Скок3".

(а) Учитати датотеку у структуру података *DataFrame*.

(б) Додати табели нову колону "Макс" и онда за сваког ученика израчунати и у ту колону уписати његов најбољи скок.

(в) Сортирати табелу по колони "Макс" и приказати првих пет редова тако сортиране табеле (да видимо ко су најбољи скакачи у разреду).

In [ ]: