11. час: Сортирање, филтрирање и фреквенцијска анализа

На овом часу ћемо говорити о:

1. преуређивању редова табеле како би се поређали по величини по неком критеријуму (*сортирање*);
2. издвајању редова табеле који задовољавају неке услове (*филтрирање*); и
3. бројању редова табеле који имају неке особине (*фреквенцијска анализа*).

11.1. Сортирање података

Сортирати податке значи поређати их по величини. Да бисмо видели како се то ради у библиотеци *pandas* прво ћемо учитати библиотеку:

In [2]:
```
# изврши ову ћелију
import pandas as pd
```

а онда ћемо поново направити табелу са подацима о групи деце коју смо већ користили, колонама ћемо дати одговарајућа имена и индексираћемо табелу именима деце:

In [3]:
```
podaci = [["Ана",     "ж", 13, 46, 160],
          ["Бојан",   "м", 14, 52, 165],
          ["Влада",   "м", 13, 47, 157],
          ["Гордана", "ж", 15, 54, 165],
          ["Дејан",   "м", 15, 56, 163],
          ["Ђорђе",   "м", 13, 45, 159],
          ["Елена",   "ж", 14, 49, 161],
          ["Жаклина", "ж", 15, 52, 164],
          ["Зоран",   "м", 15, 57, 167],
          ["Ивана",   "ж", 13, 45, 158],
          ["Јасна",   "ж", 14, 51, 162]]
tabela = pd.DataFrame(podaci)
tabela.columns=["Име", "Пол", "Старост", "Тежина", "Висина"]
tabela1 = tabela.set_index("Име")
```

Ево како табела изгледа:

In [4]:
```
# изврши ову ћелију
tabela1
```

Out[4]:

Име	Пол	Старост	Тежина	Висина
Ана	ж	13	46	160
Бојан	м	14	52	165
Влада	м	13	47	157
Гордана	ж	15	54	165
Дејан	м	15	56	163
Ђорђе	м	13	45	159
Елена	ж	14	49	161
Жаклина	ж	15	52	164
Зоран	м	15	57	167
Ивана	ж	13	45	158
Јасна	ж	14	51	162

Хајде сада да сортирамо табелу по висини употребом функције `sort_values` (енгл. *sort* значи "сортирај, поређај по величини", док *values* значи "вредности").

Овој функцији морамо да кажемо по ком критеријуму се сортирају подаци (по висини, тежини, старости, ...) тако што име одговарајуће колоне наведемо као вредност аргумента `by` (енгл. реч "by" значи свашта, али у овом контексту значи "према").

Функција не мења полазну табелу, већ од ње прави нову:

In [4]:
```
# изврши ову ћелију
tabela1_po_visini = tabela1.sort_values(by="Висина")
tabela1_po_visini
```

Out[4]:

Име	Пол	Старост	Тежина	Висина
Влада	м	13	47	157
Ивана	ж	13	45	158
Ђорђе	м	13	45	159
Ана	ж	13	46	160
Елена	ж	14	49	161
Јасна	ж	14	51	162
Дејан	м	15	56	163
Жаклина	ж	15	52	164
Бојан	м	14	52	165
Гордана	ж	15	54	165
Зоран	м	15	57	167

Пошто нисмо навели како желимо да сортирамо податке (од најмањег ка највећем, или обрнуто) подаци су сортирани од најмањег ка највећем. Уколико желимо да сортирамо табелу по висини, али од највеће ка најмањој, потребно је то нагласити користећи параметар `ascending=False` (енгл. *ascending* значи "растуће").

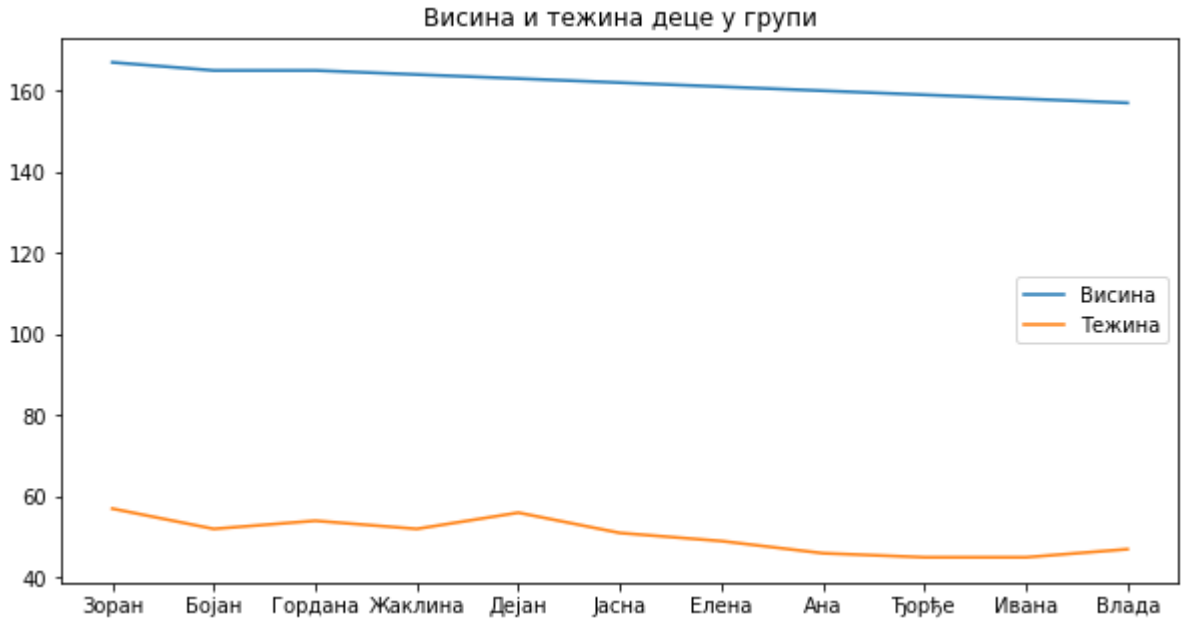
In [5]:
```
# изврши ову ћелију
tabela1_po_visini = tabela1.sort_values(by="Висина", ascending=False)
tabela1_po_visini
```

Out[5]:

Име	Пол	Старост	Тежина	Висина
Зоран	м	15	57	167
Бојан	м	14	52	165
Гордана	ж	15	54	165
Жаклина	ж	15	52	164
Дејан	м	15	56	163
Јасна	ж	14	51	162
Елена	ж	14	49	161
Ана	ж	13	46	160
Ђорђе	м	13	45	159
Ивана	ж	13	45	158
Влада	м	13	47	157

Хајде, за крај, да прикажемо податке из овако сортиране табеле.

In [9]:
```
# изврши ову ћелију
import matplotlib.pyplot as plt
plt.figure(figsize=(10,5))
plt.plot(tabela1_po_visini.index, tabela1_po_visini["Висина"], label="Висина")
plt.plot(tabela1_po_visini.index, tabela1_po_visini["Тежина"], label="Тежина")
plt.title("Висина и тежина деце у групи")
plt.legend()
plt.show()
plt.close()
```



11.2. Филтрирање података

Често је из табеле потребно издвојити редове који имају неке особине. На пример, ако желимо да издвојимо само оне редове табеле у којима су наведени подаци о девојчицама, то можемо урадити на следећи начин:

```
tabela1[tabela1.Пол == "ж"]
```

Овај израз ће из табеле `tabela1` издвојити све редове код којих у колони "Пол" пише "ж". (Обратите пажњу на то да се приликом формирања критеријума у изразу `tabela1.Пол` не пишу наводници! Не питајте зашто...)

In [6]:
```
# изврши ову ћелију
devojke = tabela1[tabela1.Пол == "ж"]
devojke
```

Out[6]:

Име	Пол	Старост	Тежина	Висина
Ана	ж	13	46	160
Гордана	ж	15	54	165
Елена	ж	14	49	161
Жаклина	ж	15	52	164
Ивана	ж	13	45	158
Јасна	ж	14	51	162

На сличан начин можемо да издвојимо сву децу која имају преко 50 кг:

In [7]:
```
# изврши ову ћелију
preko_50kg = tabela1[tabela1.Тежина > 50]
preko_50kg
```

Out[7]:

Име	Пол	Старост	Тежина	Висина
Бојан	м	14	52	165
Гордана	ж	15	54	165
Дејан	м	15	56	163
Жаклина	ж	15	52	164
Зоран	м	15	57	167
Јасна	ж	14	51	162

Критеријуме можемо и да комбинујемо. На пример, ако желимо да из табеле извучемо податке о свим дечацима са највише 55 кг, то можемо учинити овако:

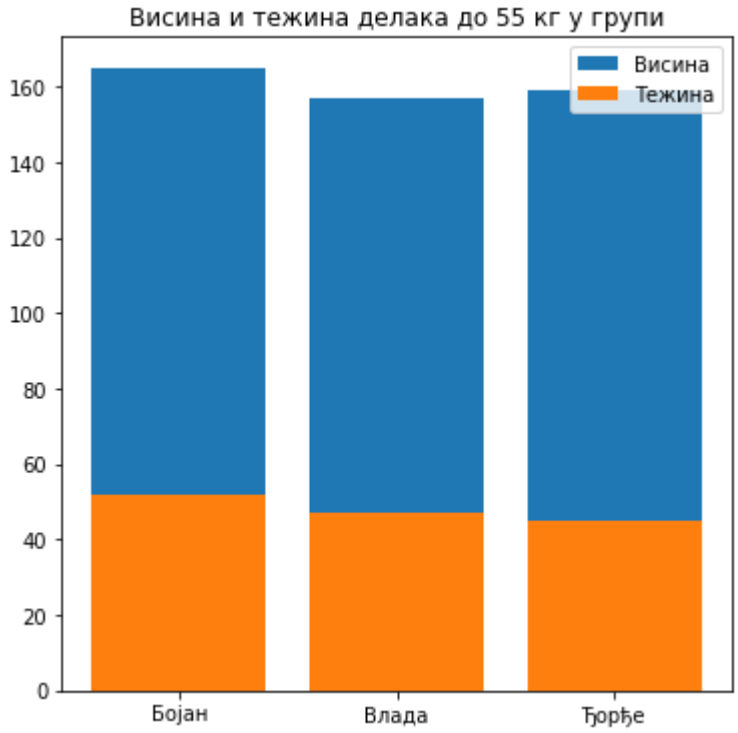
In [8]:
```
# изврши ову ћелију
decaci_do_55kg = tabela1[(tabela1.Тежина <= 55) & (tabela1.Пол == "м")]
decaci_do_55kg
```

Out[8]:

Име	Пол	Старост	Тежина	Висина
Бојан	м	14	52	165
Влада	м	13	47	157
Ђорђе	м	13	45	159

Приказаћемо, за крај, податке о тежини и висини ових дечака једним графиконом:

In [13]:
```
# изврши ову ћелију
plt.figure(figsize=(6,6))
plt.bar(decaci_do_55kg.index, decaci_do_55kg["Висина"], label="Висина")
plt.bar(decaci_do_55kg.index, decaci_do_55kg["Тежина"], label="Тежина")
plt.title("Висина и тежина делака до 55 кг у групи")
plt.legend()
plt.show()
plt.close()
```



11.3. Фреквенцијска анализа

Да се подсетимо, фреквенцијска анализа низа података се своди на то да се преброје подаци исте врсте у низу. Док смо раније морали доста тога сами да урадимо, библиотека `pandas` има функцију `value_counts` која врши фреквенцијску анализу (енгл. *value* значи "вредност", док *count* значи "бројати"; дакле, пребројати вредности).

Ево примера. Ако у табели са којом радимо желимо да пребројимо дечаке и девојчице, то можемо учинити позивом функције `value_counts` овако:

In [9]:
```
# изврши ову ћелију
tabela1["Пол"].value_counts()
```

Out[9]:
```
ж    6
м    5
Name: Пол, dtype: int64
```

Функција `value_counts` је у колони "Пол" пребројала све вредности и утврдила да се у тој колони вредност "ж" појављује 6 пута, док се вредност "м" појављује 5 пута.

Ако желимо да утврдимо старосну структуру групе, применићемо функцију `value_counts` на колону "Старост":

In [10]:
```
# изврши ову ћелију
tabela1["Старост"].value_counts()
```

Out[10]:
```
15    4
13    4
14    3
Name: Старост, dtype: int64
```

Функција `value_counts` је у колони "Старост" пребројала све вредности и утврдила да се у тој колони вредности 15 и 13 појављују по 4 пута, док се вредност 14 појављује 3 пута.

Ако резултат рада функције `value_counts` сместимо у променљиву:

In [11]:
```
# изврши ову ћелију
frekv = tabela1["Пол"].value_counts()
frekv
```

Out[11]:
```
ж    6
м    5
Name: Пол, dtype: int64
```

онда можемо лако да реконструишемо које су вредности уочене у табели, и које су њихове фреквенције:

```
frekv.index
```

нам даје листу уочених вредности, док

```
frekv.values
```

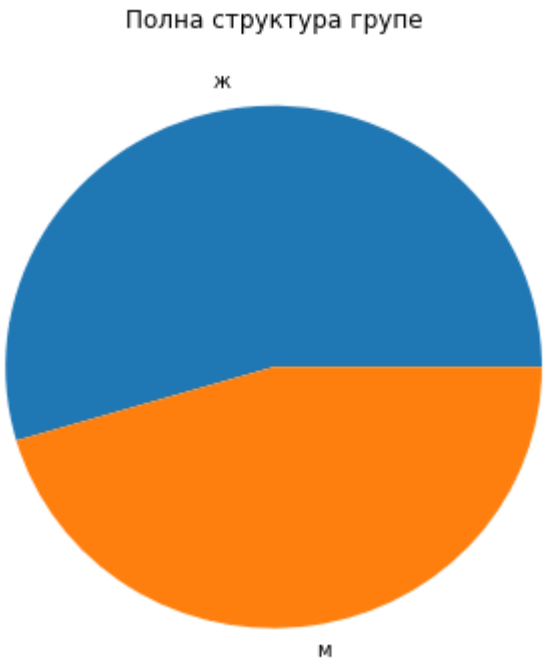
даје њихове фреквенције.

In [12]:
```
# изврши ову ћелију
print("Вредности које се јављају у колони:", frekv.index)
print("Њихове фреквенције:", frekv.values)
```

```
Вредности које се јављају у колони: Index(['ж', 'м'], dtype='object')
Њихове фреквенције: [6 5]
```

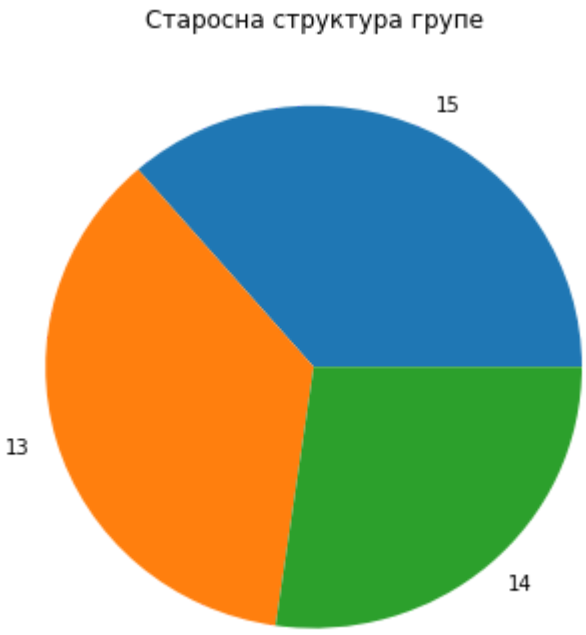
Полну структуру ове групе деце можемо да прикажемо секторским дијаграмом овако:

In [17]:
```
import matplotlib.pyplot as plt
frekv = tabela1["Пол"].value_counts()
plt.figure(figsize=(6,6))
plt.pie(frekv.values, labels=frekv.index)
plt.title("Полна структура групе")
plt.show()
plt.close()
```



На сличан начин можемо да прикажемо старосну структуру групе:

In [18]:
```
frekv = tabela1["Старост"].value_counts()
plt.figure(figsize=(6,6))
plt.pie(frekv.values, labels=frekv.index)
plt.title("Старосна структура групе")
plt.show()
plt.close()
```



11.4. Задаци

Задатак 1. У табели испод наведене су најдуже реке Србије (дужине су дате у км):

Река	Укупна дужина	Дужина тока кроз Србију
Дунав	2.850	588
Тиса	966	164
Сава	945	207
Велика Морава (са Ј. Моравом)	480	480
Тамиш	359	118
Дрина	346	220
Западна Морава	308	308

(а) Представити ове податке табеларно, а онда сортирати табелу по дужини тока реке кроз Србију.

In [ ]:
```

```

(б) За наведене реке приказати линијским дијаграмом укупну дужину, и дужину тока реке кроз

Србију.

In [ ]:

(в) Од дате табеле направити нову у којој су издвојене само оне реке које бар половину свог тока протичу кроз Србију.

In [ ]:

Задатак 2. Нутритивни подаци за неке рибе и морске плодове су дати у следећој табели:

Намирница (100г)	Енергетска вредност (kcal)	Угљени хидрати (г)	Беланчевине (г)	Масти (г)
Туна	116	0	26	1
Ослић	88	0	17.2	0.8
Пастрмка	119	0	18	5
Лосос	116	0	20	3.5
Скуша	205	0	19	14
Сардине	135	0	18	5
Харинга	158	0	18	9
Бакалар	82	0	18	0.7
Сом	95	0	16.4	2.8
Шаран	127	0	17.6	5.6
Орада	115	0	16.5	5.5
Јегуља	184	0	18.4	11.7
Шкампи	106	1	20	2
Дагње	86	4	12	2
Козице	71	1	13	1
Лигње	92	3	15.6	1.3
Хоботница	81	0	16.4	0.9
Јастог	112	0	20	1.5

Подаци из табеле су представљени листом у ћелији испод:

In [32]:
```
# изврши ову ћелију
morski_plodovi = [
  ["Туна", 116, 0, 26, 1],
  ["Ослић", 88, 0, 17.2, 0.8],
  ["Пастрмка", 119, 0, 18, 5],
  ["Лосос", 116, 0, 20, 3.5],
  ["Скуша", 205, 0, 19, 14],
  ["Сардине", 135, 0, 18, 5],
  ["Харинга", 158, 0, 18, 9],
  ["Бакалар", 82, 0, 18, 0.7],
  ["Сом", 95, 0, 16.4, 2.8],
  ["Шаран", 127, 0, 17.6, 5.6],
  ["Орада", 115, 0, 16.5, 5.5],
  ["Јегуља", 184, 0, 18.4, 11.7],
  ["Шкампи", 106, 1, 20, 2],
  ["Дагње", 86, 4, 12, 2],
  ["Козице", 71, 1, 13, 1],
  ["Лигње", 92, 3, 15.6, 1.3],
  ["Хоботница", 81, 0, 16.4, 0.9],
  ["Јастог", 112, 0, 20, 1.5]]
```

(а) Од ове листе у ћелији испод направи *DataFrame* и дај колонама табеле погодна имена. Предлажемо да свакој колони даш име које ће бити само једна реч (рецимо "Намирница", "ЕнергВр", "УХ", "Бел", "Масти") како би у каснијим задацима лакше именовао колоне табеле.

In [ ]:

(б) Соритрај табелу по енергетској вредности намирнице од највеће ка најмањој вредности и прикажи хистограмом тако сортиране енергетске вредности.

In [ ]:

(в) Од овако сортиране табеле направи нову у којој су само оне намирнице које не садрже угљене хидрате и имају мање од 10 г масти на 100 г намирнице.

In [ ]:

(г) Направи фреквенцијску анализу ових података према количини угљених хидрата и прикажи резултате анализе секторским дијаграмом.

In [ ]: