lekce

May 4, 2020

1 Agregace

Pracujeme s datasetem o výsledcích u maturity. Máme k dispozici výsledky ze 3 učeben uložené v souborech `u202.csv`, `u203.csv` a `u302.csv`.

[1]:
```
import pandas
```

[2]:
```
!cat u202.csv
```

```
jméno,předmět,známka,den
Jana Zbořilová,Chemie,,pá
Lukáš Jurčík,Dějepis,3,pá
Pavel Horák,Matematika,2,út
Lukáš Jurčík,Společenské vědy,2,pá
Pavel Kysilka,Biologie,1,pá
Kateřina Novotná,Dějepis,1,po
Marie Krejcárková,Fyzika,2,čt
Vasil Lácha,Dějepis,4,po
Alexey Opatrný,Matematika,2,po
Petr Valenta,Dějepis,,pá
Miroslav Bednář,Chemie,2,st
Pavel Horák,Chemie,5,út
Ivana Dvořáková,Matematika,1,st
Lenka Jarošová,Biologie,4,st
Miroslav Bednář,Dějepis,5,st
```

[3]:
```
u202 = pandas.read_csv("u202.csv", encoding="utf-8")
u203 = pandas.read_csv("u203.csv", encoding="utf-8")
u302 = pandas.read_csv("u302.csv", encoding="utf-8")
```

1.1 1. Práce s chybějícími hodnotami

[4]:
```
u202
```

[4]:

	jméno	předmět	známka	den
0	Jana Zbořilová	Chemie	NaN	pá
1	Lukáš Jurčík	Dějepis	3.0	pá
2	Pavel Horák	Matematika	2.0	út

```
3        Lukáš Jurčík  Společenské vědy     2.0  pá
4       Pavel Kysilka          Biologie     1.0  pá
5    Kateřina Novotná           Dějepis     1.0  po
6   Marie Krejcárková            Fyzika     2.0  čt
7         Vasil Lácha           Dějepis     4.0  po
8      Alexey Opatrný        Matematika     2.0  po
9        Petr Valenta           Dějepis     NaN  pá
10    Miroslav Bednář            Chemie     2.0  st
11        Pavel Horák            Chemie     5.0  út
12    Ivana Dvořáková        Matematika     1.0  st
13     Lenka Jarošová          Biologie     4.0  st
14    Miroslav Bednář           Dějepis     5.0  st
```

Na řádcích 0 a 9 je ve sloupci `známka` hodnota `NaN`. Tím se v Pandas značí chybějící hodnota, podobně jako `NULL` v SQL. V samotném CSV souboru údaj chybí úplně, mezi sousedními čárkami nic není.

Všimněte si, že ačkoli je známka celé číslo, sloupec `známka` je uložen jako desetinná čísla. Je to proto, že hodnotu `NaN` není možné reprezentovat v celých číslech. V desetinných (`float`) je pro to naopak vyhrazená hodnota. Podobná situace nastane v případě chybějících pravdivostních hodnot.

Chybějící hodnoty mohou vzniknout různými způsoby, např. * Porucha měřícího přístroje; * Chyba při načítání souboru; * Hodnota v daném kontextu nedává smysl; * Nenastane událost, na kterou čekáme; * …

Často taky nemáme tušení a můžeme jen hádat.

V různých situacích dává smysl vypořádat se s chybějícími hodnotami různými způsoby.

1.1.1 Jak zjistit které hodnoty chybí?

[5]:
```
u202["známka"].isnull()
```

[5]:
```
0      True
1     False
2     False
3     False
4     False
5     False
6     False
7     False
8     False
9      True
10    False
11    False
12    False
13    False
14    False
Name: známka, dtype: bool
```

[6]:
```
u202.loc[u202["známka"].isnull()]
```

[6]:

	jméno	předmět	známka	den
0	Jana Zbořilová	Chemie	NaN	pá
9	Petr Valenta	Dějepis	NaN	pá

1.1.2 Počet, procento chybějících hodnot

[7]:
```
False + False, False + True, True + True
```

[7]: (0, 1, 2)

[8]:
```
u202["známka"].isnull().sum()
```

[8]: 2

[9]:
```
u202.shape
```

[9]: (15, 4)

[10]:
```
pocet_radku = u202.shape[0]
relativni_cetnost = u202["známka"].isnull().sum() / pocet_radku
print(f"Chybí {relativni_cetnost * 100} % hodnot.")
```

```
Chybí 13.333333333333334 % hodnot.
```

[11]:
```
relativni_cetnost
```

[11]: 0.13333333333333333

[12]:
```
u202["známka"].isnull().mean()
```

[12]: 0.13333333333333333

Side note: ten formát co ukazuje 8347359378947389347 desetinných míst není moc čitelný.

[13]:
```
print(f"Chybí {relativni_cetnost * 100:.02f} % hodnot.")
```

```
Chybí 13.33 % hodnot.
```

Mnohem lepší.

1.1.3 Jak zahodit řádky s chybějícími hodnotami?

[14]:
```
u202
```

[14]:

	jméno	předmět	známka	den
0	Jana Zbořilová	Chemie	NaN	pá

```
1        Lukáš Jurčík           Dějepis     3.0  pá
2         Pavel Horák        Matematika     2.0  út
3        Lukáš Jurčík  Společenské vědy     2.0  pá
4       Pavel Kysilka          Biologie     1.0  pá
5    Kateřina Novotná           Dějepis     1.0  po
6   Marie Krejcárková            Fyzika     2.0  čt
7         Vasil Lácha           Dějepis     4.0  po
8      Alexey Opatrný        Matematika     2.0  po
9        Petr Valenta           Dějepis     NaN  pá
10    Miroslav Bednář            Chemie     2.0  st
11        Pavel Horák            Chemie     5.0  út
12    Ivana Dvořáková        Matematika     1.0  st
13     Lenka Jarošová          Biologie     4.0  st
14    Miroslav Bednář           Dějepis     5.0  st
```

[15]:
```
u202.dropna()
```

[15]:
```
                jméno           předmět  známka den
1        Lukáš Jurčík           Dějepis     3.0  pá
2         Pavel Horák        Matematika     2.0  út
3        Lukáš Jurčík  Společenské vědy     2.0  pá
4       Pavel Kysilka          Biologie     1.0  pá
5    Kateřina Novotná           Dějepis     1.0  po
6   Marie Krejcárková            Fyzika     2.0  čt
7         Vasil Lácha           Dějepis     4.0  po
8      Alexey Opatrný        Matematika     2.0  po
10    Miroslav Bednář            Chemie     2.0  st
11        Pavel Horák            Chemie     5.0  út
12    Ivana Dvořáková        Matematika     1.0  st
13     Lenka Jarošová          Biologie     4.0  st
14    Miroslav Bednář           Dějepis     5.0  st
```

Lze přizpůsobit dalšími argumenty - minimální počet chybějících hodnot, uvažovat jen vybrané sloupce, ...

1.1.4 Jak zahodit sloupce s chybějícími hodnotami?

[16]:
```
u202
```

[16]:
```
                jméno           předmět  známka den
0      Jana Zbořilová            Chemie     NaN  pá
1        Lukáš Jurčík           Dějepis     3.0  pá
2         Pavel Horák        Matematika     2.0  út
3        Lukáš Jurčík  Společenské vědy     2.0  pá
4       Pavel Kysilka          Biologie     1.0  pá
5    Kateřina Novotná           Dějepis     1.0  po
6   Marie Krejcárková            Fyzika     2.0  čt
```

```
7         Vasil Lácha           Dějepis     4.0  po
8       Alexey Opatrný        Matematika     2.0  po
9         Petr Valenta           Dějepis     NaN  pá
10     Miroslav Bednář            Chemie     2.0  st
11         Pavel Horák            Chemie     5.0  út
12     Ivana Dvořáková        Matematika     1.0  st
13      Lenka Jarošová          Biologie     4.0  st
14     Miroslav Bednář           Dějepis     5.0  st
```

[17]:
```
u202.dropna(axis="columns")
```

[17]:
```
                jméno           předmět den
0      Jana Zbořilová            Chemie  pá
1        Lukáš Jurčík           Dějepis  pá
2         Pavel Horák        Matematika  út
3        Lukáš Jurčík  Společenské vědy  pá
4       Pavel Kysilka          Biologie  pá
5     Kateřina Novotná           Dějepis  po
6    Marie Krejcárková            Fyzika  čt
7         Vasil Lácha           Dějepis  po
8       Alexey Opatrný        Matematika  po
9         Petr Valenta           Dějepis  pá
10     Miroslav Bednář            Chemie  st
11        Pavel Horák            Chemie  út
12     Ivana Dvořáková        Matematika  st
13      Lenka Jarošová          Biologie  st
14     Miroslav Bednář           Dějepis  st
```

1.1.5 Jak nahradit chybějící hodnoty něčím jiným?

[18]:
```
u202
```

[18]:
```
                jméno           předmět  známka den
0      Jana Zbořilová            Chemie     NaN  pá
1        Lukáš Jurčík           Dějepis     3.0  pá
2         Pavel Horák        Matematika     2.0  út
3        Lukáš Jurčík  Společenské vědy     2.0  pá
4       Pavel Kysilka          Biologie     1.0  pá
5     Kateřina Novotná           Dějepis     1.0  po
6    Marie Krejcárková            Fyzika     2.0  čt
7         Vasil Lácha           Dějepis     4.0  po
8       Alexey Opatrný        Matematika     2.0  po
9         Petr Valenta           Dějepis     NaN  pá
10     Miroslav Bednář            Chemie     2.0  st
11        Pavel Horák            Chemie     5.0  út
12     Ivana Dvořáková        Matematika     1.0  st
13      Lenka Jarošová          Biologie     4.0  st
```

```
14    Miroslav Bednář           Dějepis     5.0  st
```

[19]:
```
doplneno = u202.fillna(5)
doplneno
```

[19]:
```
                jméno           předmět  známka den
0      Jana Zbořilová            Chemie     5.0  pá
1        Lukáš Jurčík           Dějepis     3.0  pá
2         Pavel Horák        Matematika     2.0  út
3        Lukáš Jurčík  Společenské vědy     2.0  pá
4       Pavel Kysilka          Biologie     1.0  pá
5    Kateřina Novotná           Dějepis     1.0  po
6   Marie Krejcárková            Fyzika     2.0  čt
7         Vasil Lácha           Dějepis     4.0  po
8      Alexey Opatrný        Matematika     2.0  po
9        Petr Valenta           Dějepis     5.0  pá
10    Miroslav Bednář            Chemie     2.0  st
11        Pavel Horák            Chemie     5.0  út
12    Ivana Dvořáková        Matematika     1.0  st
13     Lenka Jarošová          Biologie     4.0  st
14    Miroslav Bednář           Dějepis     5.0  st
```

[20]:
```
doplneno["známka"] = doplneno["známka"].astype(int)
doplneno
```

[20]:
```
                jméno           předmět  známka den
0      Jana Zbořilová            Chemie       5  pá
1        Lukáš Jurčík           Dějepis       3  pá
2         Pavel Horák        Matematika       2  út
3        Lukáš Jurčík  Společenské vědy       2  pá
4       Pavel Kysilka          Biologie       1  pá
5    Kateřina Novotná           Dějepis       1  po
6   Marie Krejcárková            Fyzika       2  čt
7         Vasil Lácha           Dějepis       4  po
8      Alexey Opatrný        Matematika       2  po
9        Petr Valenta           Dějepis       5  pá
10    Miroslav Bednář            Chemie       2  st
11        Pavel Horák            Chemie       5  út
12    Ivana Dvořáková        Matematika       1  st
13     Lenka Jarošová          Biologie       4  st
14    Miroslav Bednář           Dějepis       5  st
```

1.1.6 Někdy to ani není třeba

- Může se stát, že se sloupci, ve kterých chybí hodnoty, ani nepotřebujeme pracovat. Bylo by pak zbytečné zahazovat nějaké řádky, protože bychom tím přišli o data.
- Pandas funkce umí s chybějícími hodnotami pracovat.

[21]:
```
u202["známka"]
```

[21]:
```
0     NaN
1     3.0
2     2.0
3     2.0
4     1.0
5     1.0
6     2.0
7     4.0
8     2.0
9     NaN
10    2.0
11    5.0
12    1.0
13    4.0
14    5.0
Name: známka, dtype: float64
```

[22]:
```
u202["známka"].mean()
```

[22]:
```
2.6153846153846154
```

1.2 2. Spojení dat

Máme k dispozici údaje o maturitách ve všech 3 místnostech, chtěli bychom je spojit dohromady do jednoho dataframe.

Nejprve zahodíme studenty, kteří nedorazili ke zkoušce.

[23]:
```
u202.dropna(inplace=True)
u203.dropna(inplace=True)
u302.dropna(inplace=True)
```

[24]:
```
maturita = pandas.concat([u202, u203, u302])
maturita.head(16)
```

[24]:

	jméno	předmět	známka	den
1	Lukáš Jurčík	Dějepis	3.0	pá
2	Pavel Horák	Matematika	2.0	út
3	Lukáš Jurčík	Společenské vědy	2.0	pá
4	Pavel Kysilka	Biologie	1.0	pá
5	Kateřina Novotná	Dějepis	1.0	po
6	Marie Krejcárková	Fyzika	2.0	čt
7	Vasil Lácha	Dějepis	4.0	po
8	Alexey Opatrný	Matematika	2.0	po
10	Miroslav Bednář	Chemie	2.0	st
11	Pavel Horák	Chemie	5.0	út

```
12    Ivana Dvořáková        Matematika     1.0  st
13     Lenka Jarošová          Biologie     4.0  st
14    Miroslav Bednář           Dějepis     5.0  st
0    Kateřina Novotná  Společenské vědy     3.0  po
1          Arnošt Sas        Matematika     5.0  po
2         Vasil Lácha       Informatika     3.0  po
```

Pandas napojilo indexy jednotlivých dataframes dohromady, takže jsou tam duplicity.

[25]:
```
maturita = pandas.concat([u202, u203, u302], ignore_index=True)
maturita.head(16)
```

[25]:
```
                jméno           předmět  známka den
0        Lukáš Jurčík           Dějepis     3.0  pá
1         Pavel Horák        Matematika     2.0  út
2        Lukáš Jurčík  Společenské vědy     2.0  pá
3       Pavel Kysilka          Biologie     1.0  pá
4    Kateřina Novotná           Dějepis     1.0  po
5   Marie Krejcárková            Fyzika     2.0  čt
6         Vasil Lácha           Dějepis     4.0  po
7      Alexey Opatrný        Matematika     2.0  po
8     Miroslav Bednář            Chemie     2.0  st
9         Pavel Horák            Chemie     5.0  út
10    Ivana Dvořáková        Matematika     1.0  st
11     Lenka Jarošová          Biologie     4.0  st
12    Miroslav Bednář           Dějepis     5.0  st
13   Kateřina Novotná  Společenské vědy     3.0  po
14         Arnošt Sas        Matematika     5.0  po
15        Vasil Lácha       Informatika     3.0  po
```

Ještě budeme chtít uchovat informaci o místnostech, ve kterých zkoušky probíhaly.

[26]:
```
u202["místnost"] = "u202"
u203["místnost"] = "u203"
u302["místnost"] = "u302"
```

[27]:
```
maturita = pandas.concat([u202, u203, u302], ignore_index=True)
maturita.head(16)
```

[27]:
```
                jméno           předmět  známka den místnost
0        Lukáš Jurčík           Dějepis     3.0  pá     u202
1         Pavel Horák        Matematika     2.0  út     u202
2        Lukáš Jurčík  Společenské vědy     2.0  pá     u202
3       Pavel Kysilka          Biologie     1.0  pá     u202
4    Kateřina Novotná           Dějepis     1.0  po     u202
5   Marie Krejcárková            Fyzika     2.0  čt     u202
6         Vasil Lácha           Dějepis     4.0  po     u202
7      Alexey Opatrný        Matematika     2.0  po     u202
```

```
8     Miroslav Bednář            Chemie     2.0  st     u202
9         Pavel Horák            Chemie     5.0  út     u202
10    Ivana Dvořáková        Matematika     1.0  st     u202
11     Lenka Jarošová          Biologie     4.0  st     u202
12    Miroslav Bednář           Dějepis     5.0  st     u202
13   Kateřina Novotná  Společenské vědy     3.0  po     u203
14         Arnošt Sas        Matematika     5.0  po     u203
15        Vasil Lácha       Informatika     3.0  po     u203
```

Výsledný dataframe uložíme do CSV souboru.

[28]:
```
maturita.to_csv("maturita.csv", index=False)
```

[29]:
```
!cat maturita.csv
```

```
jméno,předmět,známka,den,místnost
Lukáš Jurčík,Dějepis,3.0,pá,u202
Pavel Horák,Matematika,2.0,út,u202
Lukáš Jurčík,Společenské vědy,2.0,pá,u202
Pavel Kysilka,Biologie,1.0,pá,u202
Kateřina Novotná,Dějepis,1.0,po,u202
Marie Krejcárková,Fyzika,2.0,čt,u202
Vasil Lácha,Dějepis,4.0,po,u202
Alexey Opatrný,Matematika,2.0,po,u202
Miroslav Bednář,Chemie,2.0,st,u202
Pavel Horák,Chemie,5.0,út,u202
Ivana Dvořáková,Matematika,1.0,st,u202
Lenka Jarošová,Biologie,4.0,st,u202
Miroslav Bednář,Dějepis,5.0,st,u202
Kateřina Novotná,Společenské vědy,3.0,po,u203
Arnošt Sas,Matematika,5.0,po,u203
Vasil Lácha,Informatika,3.0,po,u203
Lenka Jarošová,Fyzika,3.0,st,u203
Marie Kortusová,Fyzika,3.0,st,u203
Monika Dudysová,Chemie,3.0,pá,u203
Josef Vodseďálek,Informatika,2.0,út,u203
Alexey Opatrný,Zeměpis,1.0,po,u203
Antonín Hlídek,Fyzika,4.0,po,u203
Monika Dudysová,Společenské vědy,3.0,pá,u203
Filip Lacina,Fyzika,2.0,čt,u203
Marta Kinclová,Biologie,4.0,st,u203
Martina Korbářová,Zeměpis,3.0,út,u203
Petr Tábor,Informatika,1.0,po,u302
Petr Tábor,Společenské vědy,1.0,po,u302
Marie Krejcárková,Matematika,5.0,čt,u302
Pavel Kysilka,Informatika,1.0,pá,u302
Josef Vodseďálek,Biologie,2.0,út,u302
Filip Lacina,Matematika,1.0,čt,u302
```

```
Marta Kinclová,Informatika,4.0,st,u302
Ivana Dvořáková,Chemie,5.0,st,u302
Marie Kortusová,Dějepis,3.0,st,u302
Arnošt Sas,Chemie,4.0,po,u302
Martina Korbářová,Informatika,3.0,út,u302
Antonín Hlídek,Matematika,3.0,po,u302
```

1.3 3. Joinování dat

Obdobou SQL příkazu `JOIN` je v Pandas funkce `merge`. K datasetu výsledků u maturitních zkoušek budeme joinovat data o předsedajících maturitních komisí v jednotlivých dnech.

[30]:
```
!cat predsedajici.csv
```

```
den,datum,jméno
po,20.5.2019,Marie Zuzaňáková
út,21.5.2019,Marie Zuzaňáková
st,22.5.2019,Petr Ortinský
čt,23.5.2019,Petr Ortinský
pá,24.5.2019,Alena Pniáčková
```

[31]:
```
preds = pandas.read_csv("predsedajici.csv", encoding="utf-8")
preds
```

[31]:

	den	datum	jméno
0	po	20.5.2019	Marie Zuzaňáková
1	út	21.5.2019	Marie Zuzaňáková
2	st	22.5.2019	Petr Ortinský
3	čt	23.5.2019	Petr Ortinský
4	pá	24.5.2019	Alena Pniáčková

Vyzkoušíme merge, zatím jen na místnosti u202.

[32]:
```
u202
```

[32]:

	jméno	předmět	známka	den	místnost
1	Lukáš Jurčík	Dějepis	3.0	pá	u202
2	Pavel Horák	Matematika	2.0	út	u202
3	Lukáš Jurčík	Společenské vědy	2.0	pá	u202
4	Pavel Kysilka	Biologie	1.0	pá	u202
5	Kateřina Novotná	Dějepis	1.0	po	u202
6	Marie Krejcárková	Fyzika	2.0	čt	u202
7	Vasil Lácha	Dějepis	4.0	po	u202
8	Alexey Opatrný	Matematika	2.0	po	u202
10	Miroslav Bednář	Chemie	2.0	st	u202
11	Pavel Horák	Chemie	5.0	út	u202
12	Ivana Dvořáková	Matematika	1.0	st	u202
13	Lenka Jarošová	Biologie	4.0	st	u202

```
14    Miroslav Bednář           Dějepis     5.0  st     u202
```

[33]:
```
test = pandas.merge(u202, preds)
test
```

[33]:
```
Empty DataFrame
Columns: [jméno, předmět, známka, den, místnost, datum]
Index: []
```

Dostali jsme prázdný dataframe. To je proto, že defaultně `merge` dělá `INNER JOIN` přes všechny sloupce se stejnými jmény, zde `jméno` a `den`. Protože jedno `jméno` odpovídá studentovi a druhé předsedovi, nemáme žádný průnik.

Nabízel by se `OUTER JOIN`, ale ten nepomůže.

[34]:
```
test = pandas.merge(u202, preds, how="outer")
test
```

[34]:

	jméno	předmět	známka	den	místnost	datum
0	Lukáš Jurčík	Dějepis	3.0	pá	u202	NaN
1	Lukáš Jurčík	Společenské vědy	2.0	pá	u202	NaN
2	Pavel Horák	Matematika	2.0	út	u202	NaN
3	Pavel Horák	Chemie	5.0	út	u202	NaN
4	Pavel Kysilka	Biologie	1.0	pá	u202	NaN
5	Kateřina Novotná	Dějepis	1.0	po	u202	NaN
6	Marie Krejcárková	Fyzika	2.0	čt	u202	NaN
7	Vasil Lácha	Dějepis	4.0	po	u202	NaN
8	Alexey Opatrný	Matematika	2.0	po	u202	NaN
9	Miroslav Bednář	Chemie	2.0	st	u202	NaN
10	Miroslav Bednář	Dějepis	5.0	st	u202	NaN
11	Ivana Dvořáková	Matematika	1.0	st	u202	NaN
12	Lenka Jarošová	Biologie	4.0	st	u202	NaN
13	Marie Zuzaňáková	NaN	NaN	po	NaN	20.5.2019
14	Marie Zuzaňáková	NaN	NaN	út	NaN	21.5.2019
15	Petr Ortinský	NaN	NaN	st	NaN	22.5.2019
16	Petr Ortinský	NaN	NaN	čt	NaN	23.5.2019
17	Alena Pniáčková	NaN	NaN	pá	NaN	24.5.2019

Tím se nám akorát promíchala jména studentů a předsedajících, navíc vznikla spousta nedefinovaných hodnot.

Ve skutečnosti potřebujeme provést `JOIN` jen podle sloupce `den` – ke každému dni známe předsedu komise a všechny studenty, kteří měli ten den zkoušku.

[35]:
```
test = pandas.merge(u202, preds, on="den")
test
```

[35]:
```
              jméno_x           předmět  známka den místnost      datum  \
0        Lukáš Jurčík           Dějepis     3.0  pá     u202  24.5.2019
1        Lukáš Jurčík  Společenské vědy     2.0  pá     u202  24.5.2019
2       Pavel Kysilka          Biologie     1.0  pá     u202  24.5.2019
3         Pavel Horák        Matematika     2.0  út     u202  21.5.2019
4         Pavel Horák             Chemie     5.0  út     u202  21.5.2019
5    Kateřina Novotná           Dějepis     1.0  po     u202  20.5.2019
6         Vasil Lácha           Dějepis     4.0  po     u202  20.5.2019
7      Alexey Opatrný        Matematika     2.0  po     u202  20.5.2019
8   Marie Krejcárková             Fyzika     2.0  čt     u202  23.5.2019
9     Miroslav Bednář             Chemie     2.0  st     u202  22.5.2019
10    Ivana Dvořáková        Matematika     1.0  st     u202  22.5.2019
11     Lenka Jarošová          Biologie     4.0  st     u202  22.5.2019
12    Miroslav Bednář           Dějepis     5.0  st     u202  22.5.2019

              jméno_y
0    Alena Pniáčková
1    Alena Pniáčková
2    Alena Pniáčková
3   Marie Zuzaňáková
4   Marie Zuzaňáková
5   Marie Zuzaňáková
6   Marie Zuzaňáková
7   Marie Zuzaňáková
8      Petr Ortinský
9      Petr Ortinský
10     Petr Ortinský
11     Petr Ortinský
12     Petr Ortinský
```

Skoro dobré, jen potřebujeme rozumně přejmenovat sloupce se jmény na jméno studenta a jméno předsedy.

[36]:
```
test = test.rename(columns={"jméno_x": "jméno_student", "jméno_y":␣
↪"jméno_předseda"})
test
```

[36]:
```
         jméno_student           předmět  známka den místnost      datum  \
0         Lukáš Jurčík           Dějepis     3.0  pá     u202  24.5.2019
1         Lukáš Jurčík  Společenské vědy     2.0  pá     u202  24.5.2019
2        Pavel Kysilka          Biologie     1.0  pá     u202  24.5.2019
3          Pavel Horák        Matematika     2.0  út     u202  21.5.2019
4          Pavel Horák             Chemie     5.0  út     u202  21.5.2019
5     Kateřina Novotná           Dějepis     1.0  po     u202  20.5.2019
6          Vasil Lácha           Dějepis     4.0  po     u202  20.5.2019
7       Alexey Opatrný        Matematika     2.0  po     u202  20.5.2019
8    Marie Krejcárková             Fyzika     2.0  čt     u202  23.5.2019
```

```
9     Miroslav Bednář            Chemie     2.0  st     u202  22.5.2019
10    Ivana Dvořáková        Matematika     1.0  st     u202  22.5.2019
11     Lenka Jarošová          Biologie     4.0  st     u202  22.5.2019
12    Miroslav Bednář           Dějepis     5.0  st     u202  22.5.2019

      jméno_předseda
0    Alena Pniáčková
1    Alena Pniáčková
2    Alena Pniáčková
3   Marie Zuzaňáková
4   Marie Zuzaňáková
5   Marie Zuzaňáková
6   Marie Zuzaňáková
7   Marie Zuzaňáková
8      Petr Ortinský
9      Petr Ortinský
10     Petr Ortinský
11     Petr Ortinský
12     Petr Ortinský
```

To už vypadá dobře, provedeme tedy `JOIN` pro celý dataset a výsledek opět uložíme do CSV.

[37]:
```
maturita2 = pandas.merge(maturita, preds, on="den")
maturita2 = maturita2.rename(columns={"jméno_x": "jméno_student", "jméno_y":␣
 ↪"jméno_předseda"})
maturita2.head()
```

[37]:
```
     jméno_student           předmět  známka den místnost      datum  \
0     Lukáš Jurčík           Dějepis     3.0  pá     u202  24.5.2019
1     Lukáš Jurčík  Společenské vědy     2.0  pá     u202  24.5.2019
2    Pavel Kysilka          Biologie     1.0  pá     u202  24.5.2019
3  Monika Dudysová            Chemie     3.0  pá     u203  24.5.2019
4  Monika Dudysová  Společenské vědy     3.0  pá     u203  24.5.2019

    jméno_předseda
0  Alena Pniáčková
1  Alena Pniáčková
2  Alena Pniáčková
3  Alena Pniáčková
4  Alena Pniáčková
```

[38]:
```
maturita2.to_csv("maturita2.csv", index=False)
```

1.4 4. Grupování

Pandí obdoba databázové operace `GROUP BY` se nazývá… `groupby`.

[39]:
```
maturita2.head()
```

[39]:

	jméno_student	předmět	známka	den	místnost	datum	jméno_předseda
0	Lukáš Jurčík	Dějepis	3.0	pá	u202	24.5.2019	Alena Pniáčková
1	Lukáš Jurčík	Společenské vědy	2.0	pá	u202	24.5.2019	Alena Pniáčková
2	Pavel Kysilka	Biologie	1.0	pá	u202	24.5.2019	Alena Pniáčková
3	Monika Dudysová	Chemie	3.0	pá	u203	24.5.2019	Alena Pniáčková
4	Monika Dudysová	Společenské vědy	3.0	pá	u203	24.5.2019	Alena Pniáčková

[40]:
```
groups = maturita2.groupby("místnost")
```

Jak to vypadá?

[41]:
```
groups
```

[41]:
```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x7f061dc17c10>
```

Wtf

[42]:
```
for key, item in groups:
    print(key)
    display(item)
```

u202

	jméno_student	předmět	známka	den	místnost	datum	jméno_předseda
0	Lukáš Jurčík	Dějepis	3.0	pá	u202	24.5.2019	Alena Pniáčková
1	Lukáš Jurčík	Společenské vědy	2.0	pá	u202	24.5.2019	Alena Pniáčková
2	Pavel Kysilka	Biologie	1.0	pá	u202	24.5.2019	Alena Pniáčková
6	Pavel Horák	Matematika	2.0	út	u202	21.5.2019	Alena Pniáčková
7	Pavel Horák	Chemie	5.0	út	u202	21.5.2019	Alena Pniáčková
12	Kateřina Novotná	Dějepis	1.0	po	u202	20.5.2019	Alena Pniáčková
13	Vasil Lácha	Dějepis	4.0	po	u202	20.5.2019	Alena Pniáčková
14	Alexey Opatrný	Matematika	2.0	po	u202	20.5.2019	Alena Pniáčková
24	Marie Krejcárková	Fyzika	2.0	čt	u202	23.5.2019	Alena Pniáčková
28	Miroslav Bednář	Chemie	2.0	st	u202	22.5.2019	Alena Pniáčková
29	Ivana Dvořáková	Matematika	1.0	st	u202	22.5.2019	Alena Pniáčková
30	Lenka Jarošová	Biologie	4.0	st	u202	22.5.2019	Alena Pniáčková
31	Miroslav Bednář	Dějepis	5.0	st	u202	22.5.2019	Alena Pniáčková

```
1    Alena Pniáčková
2    Alena Pniáčková
6   Marie Zuzaňáková
7   Marie Zuzaňáková
12  Marie Zuzaňáková
13  Marie Zuzaňáková
14  Marie Zuzaňáková
24     Petr Ortinský
28     Petr Ortinský
29     Petr Ortinský
30     Petr Ortinský
31     Petr Ortinský
```

u203

```
        jméno_student           předmět  známka den místnost      datum  \
3     Monika Dudysová            Chemie     3.0  pá     u203  24.5.2019
4     Monika Dudysová  Společenské vědy     3.0  pá     u203  24.5.2019
8    Josef Vodseďálek       Informatika     2.0  út     u203  21.5.2019
9   Martina Korbářová           Zeměpis     3.0  út     u203  21.5.2019
15   Kateřina Novotná  Společenské vědy     3.0  po     u203  20.5.2019
16         Arnošt Sas        Matematika     5.0  po     u203  20.5.2019
17        Vasil Lácha       Informatika     3.0  po     u203  20.5.2019
18     Alexey Opatrný           Zeměpis     1.0  po     u203  20.5.2019
19     Antonín Hlídek            Fyzika     4.0  po     u203  20.5.2019
25       Filip Lacina            Fyzika     2.0  čt     u203  23.5.2019
32     Lenka Jarošová            Fyzika     3.0  st     u203  22.5.2019
33    Marie Kortusová            Fyzika     3.0  st     u203  22.5.2019
34     Marta Kinclová          Biologie     4.0  st     u203  22.5.2019

      jméno_předseda
3    Alena Pniáčková
4    Alena Pniáčková
8   Marie Zuzaňáková
9   Marie Zuzaňáková
15  Marie Zuzaňáková
16  Marie Zuzaňáková
17  Marie Zuzaňáková
18  Marie Zuzaňáková
19  Marie Zuzaňáková
25     Petr Ortinský
32     Petr Ortinský
33     Petr Ortinský
34     Petr Ortinský
```

u302

```
        jméno_student           předmět  známka den místnost      datum  \
```

```
5       Pavel Kysilka       Informatika     1.0  pá     u302  24.5.2019
10   Josef Vodseďálek          Biologie     2.0  út     u302  21.5.2019
11  Martina Korbářová       Informatika     3.0  út     u302  21.5.2019
20         Petr Tábor       Informatika     1.0  po     u302  20.5.2019
21         Petr Tábor  Společenské vědy     1.0  po     u302  20.5.2019
22         Arnošt Sas            Chemie     4.0  po     u302  20.5.2019
23     Antonín Hlídek        Matematika     3.0  po     u302  20.5.2019
26  Marie Krejcárková        Matematika     5.0  čt     u302  23.5.2019
27       Filip Lacina        Matematika     1.0  čt     u302  23.5.2019
35     Marta Kinclová       Informatika     4.0  st     u302  22.5.2019
36    Ivana Dvořáková            Chemie     5.0  st     u302  22.5.2019
37    Marie Kortusová           Dějepis     3.0  st     u302  22.5.2019

      jméno_předseda
5    Alena Pniáčková
10  Marie Zuzaňáková
11  Marie Zuzaňáková
20  Marie Zuzaňáková
21  Marie Zuzaňáková
22  Marie Zuzaňáková
23  Marie Zuzaňáková
26     Petr Ortinský
27     Petr Ortinský
35     Petr Ortinský
36     Petr Ortinský
37     Petr Ortinský
```

Aha!

[43]:
```
groups.count()
```

[43]:
```
           jméno_student  předmět  známka  den  datum  jméno_předseda
místnost
u202                  13       13      13   13     13              13
u203                  13       13      13   13     13              13
u302                  12       12      12   12     12              12
```

Můžeme si spočítat průměrnou známku přes předměty.

Jako series.

[44]:
```
maturita2.groupby('předmět')['známka'].mean()
```

[44]:
```
předmět
Biologie    2.750000
Chemie      3.800000
Dějepis     3.200000
Fyzika      2.800000
```

```
Informatika          2.333333
Matematika           2.714286
Společenské vědy     2.250000
Zeměpis              2.000000
Name: známka, dtype: float64
```

Jako dataframe.

[45]:
```
maturita2.groupby('předmět')[['známka']].mean()
```

[45]:
```
                     známka
předmět
Biologie           2.750000
Chemie             3.800000
Dějepis            3.200000
Fyzika             2.800000
Informatika        2.333333
Matematika         2.714286
Společenské vědy   2.250000
Zeměpis            2.000000
```

Pandy nejsou blbé a průměry počítají jen z numerických sloupců, což je v našem případě jen `známka`.

[46]:
```
maturita2.groupby('předmět').mean()
```

[46]:
```
                     známka
předmět
Biologie           2.750000
Chemie             3.800000
Dějepis            3.200000
Fyzika             2.800000
Informatika        2.333333
Matematika         2.714286
Společenské vědy   2.250000
Zeměpis            2.000000
```

Další agregační funkce jsou např. * `sum` - součet hodnot * `max` - maximální hodnota * `min` - minimální hodnota * `first` - první hodnota * `last` - poslední hodnota * `mean` - průměr z hodnot * `median` - medián z hodnot * `var` - rozptyl hodnot * `std` - standardní odchylka hodnot * `all` - True, pokud jsou všechny hodnoty True * `any` - True, pokud je alespoň jedna z hodnot True

Můžeme se podívat, jak vypadaly známky mezi jednotlivými předsedy komise.

[47]:
```
groups_preds = maturita2.groupby("jméno_předseda")
```

[48]:
```
groups_preds.mean()
```

[48]:
```
                     známka
jméno_předseda
Alena Pniáčková    2.166667
Marie Zuzaňáková   2.722222
Petr Ortinský      3.142857
```

[49]:
```
groups_preds.median()
```

[49]:
```
                   známka
jméno_předseda
Alena Pniáčková       2.5
Marie Zuzaňáková      3.0
Petr Ortinský         3.0
```

[50]:
```
groups_preds.std()
```

[50]:
```
                     známka
jméno_předseda
Alena Pniáčková    0.983192
Marie Zuzaňáková   1.319784
Petr Ortinský      1.406422
```

Když chceme víc agregací najednou, můžeme použít funkci `agg` a dát jí seznam názvů funkcí, které chceme aplikovat.

[51]:
```
groups_preds.agg(["mean", "median", "std"])
```

[51]:
```
                     známka
                       mean median       std
jméno_předseda
Alena Pniáčková    2.166667    2.5  0.983192
Marie Zuzaňáková   2.722222    3.0  1.319784
Petr Ortinský      3.142857    3.0  1.406422
```

[ ]:
```

```