PRAHACODINGSCHOOL

Python Pandas

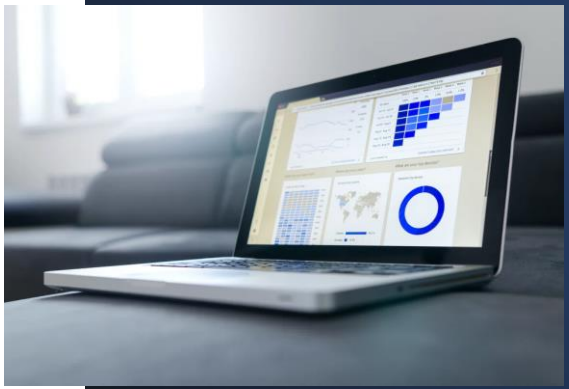
Výběr dat, filtrování

Operace s tabulkami a daty

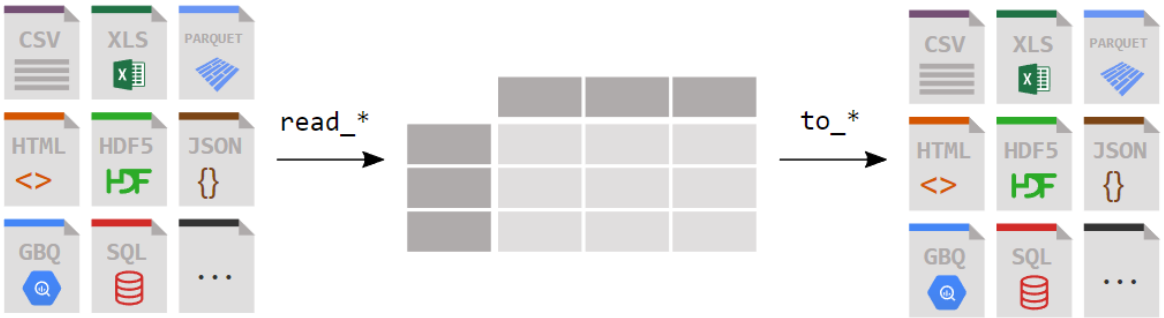
Lektor: Martin Rosický

PANDAS

svatý grál analýzy dat



Pandas: základem je DataFrame



source: https://pandas.pydata.org/

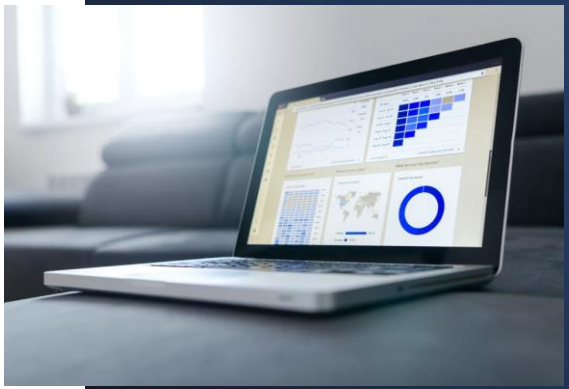
- Univerzální nástroj pro zpracování a analýzu dat
- Základ tvoří
  - pandas.Series (1D data)
  - pandas.DataFrame (2D data)
- Široká paleta nástrojů pro import/export dat
- Manipulace s daty inspirované (, ale ne omezené na) SQL
- Vyvinuto s důrazem na výkon

PANDAS
svatý grál analýzy dat

Zobrazení dat

Řazení

Filtrování



Eclipse PyDev Console

```
> import pandas                                    # import knihovny pandas
> sklad = pandas.read_csv(<vytvořený csv soubor>)  # import dat z csv
>
> sklad.info()                                     # zobrazení informace o datech
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 807 entries, 0 to 806
Data columns (total 9 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   skupina   807 non-null    object
 1   pid       807 non-null    object
 2   produkt   807 non-null    object
 3   varianta  807 non-null    object
 4   prodej    807 non-null    float64
 5   nakup     807 non-null    float64
 6   skladem   807 non-null    int64
 7   prodano   807 non-null    int64
 8   marze     807 non-null    float64
dtypes: float64(3), int64(2), object(4)
memory usage: 56.9+ KB
```

Náhled dat

```
> sklad                                                # zobrazení vzorku dat
    skupina                pid    produkt  ... skladem  prodano  marze
  0    kosmetika              19106       olej  ...      71       48  28.72
  1    kosmetika               1910       olej  ...      75       51  14.37
  2    kosmetika              19376       olej  ...      87       99  13.51
  3    kosmetika               1937       olej  ...      51       17  13.23
  4    kosmetika              19503       olej  ...      81       77  13.86
  ..         ...                ...        ...  ...     ...      ...    ...
  802     mobily  100153992486594583       sony  ...      49       84  11.91
  803     mobily  100153992486594584  microsoft  ...      91      101  25.44
  804     mobily  100153992486594585    yealink  ...      32       28  13.73
  805     mobily  100153992486594586    yealink  ...      13       96  11.28
  806     mobily  100153992486594587    yealink  ...       1       39  11.57

  [807 rows x 10 columns]
```

Výběr sloupců a řazení

```
> sklad[['varianta','prodej','nakup']]                    # výběr sloupců podle jména
                                    varianta    prodej     nakup
0          ČISTÍCÍ OLEJ NA PLEŤ 3 v 1 - 50 ml    277.69    197.95
1    ČISTÍCÍ OLEJ NA PLEŤ 3 v 1 - vzorek 13 ml     65.29     55.91
..                                        ...       ...       ...
806                        Yealink SIP-T33G   1303.31   1152.56
> sklad.iloc[9:15,3:6]                                    # výběr rozsahem řádků a sloupců
                                varianta  prodej   nakup
9                JOJOBOVÝ OLEJ LZS - 100 ml  276.03  231.88
10  ROSA FEMINNE CREME (růžový krém) - 30 ml  401.65  329.17
11               KRÉM NA DOBRÉ RÁNO - 30 ml  317.36  279.99
12                     LALIUM - MAST - 5 ml   91.74   78.15
13                    LALIUM - MAST - 50 ml  262.81  192.05
14                   LALIUM - MAST - 100 ml  462.81  398.35
```

Řazení; vícesložkové, sestupně

```
> sklad[['varianta','prodej','nakup']].sort_values('prodej')
                                        varianta    prodej     nakup
619                      RAVENSARA AROMATICA - 100 ml      1.00      0.82
433       LÉKOVKA SKLENĚNÁ HNĚDÁ 50ml - balení 105 ks      1.00      0.89
552  DELFÍNEK - DĚTSKÝ KOUPELOVÝ OLEJ velký - 1000 ml      1.00      0.71
415       LÉKOVKA SKLENĚNÁ HNĚDÁ 10ml - balení 160 ks      1.00      0.74
..                                                 ...       ...       ...
> sklad[['varianta','prodej','nakup']].sort_values(['prodej','nakup'])
                                        varianta    prodej     nakup
147                                   MANUKA - 20 ml      1.00      0.71
552  DELFÍNEK - DĚTSKÝ KOUPELOVÝ OLEJ velký - 1000 ml      1.00      0.71
192               SMRK ZTEPILÝ, jehličí, PSP - 100 ml      1.00      0.73
435       LÉKOVKA SKLENĚNÁ HNĚDÁ 20ml - balení 180 ks      1.00      0.73
..                                                 ...       ...       ...
> sklad[['varianta','prodej','nakup']].sort_values(['prodej','nakup'],ascending=False)
                                          varianta    prodej     nakup
693          Apple iPhone 14 Pro Max 1 TB Space Black  44195.87  33818.95
694              Apple iPhone 14 Pro 1 TB Space Black  41303.31  30418.62
695        Apple iPad Pro 12.9 2022 512 GB Space Grey  40476.86  34688.01
741                     Samsung Galaxy Z Fold4 512 GB  39650.41  32816.99
..                                                ...       ...       ...
```

Filtrování dat

```
> sk = sklad[['varianta','prodej','nakup']]       # ‚sk‘ je obdobou view
> sk[(sk['prodej'] >= 10) & (sk['prodej'] <= 20)]
```

	varianta	prodej	nakup
312	LÁHEV PET S UZÁVĚREM BÍLÁ 30ml	14.88	10.80
313	LÁHEV PET S UZÁVĚREM BÍLÁ 50ml	15.70	14.03
316	LÉKOVKA SKLENĚNÁ HNĚDÁ 100ml - 1 ks	18.18	12.94
319	DÓZA PLASTOVÁ BÍLÁ - 5 ml	14.88	10.56
320	DÓZA PLASTOVÁ BÍLÁ - 30 ml	18.18	14.60
321	DÓZA PLASTOVÁ BÍLÁ - 50 ml	19.83	17.75
432	LÉKOVKA SKLENĚNÁ HNĚDÁ 50ml - 1 ks	14.88	12.94
..	...	...	...
547	KAPACÍ SKLENĚNÁ PIPETA na lékovku 10ml	16.53	11.94
548	KAPACÍ SKLENĚNÁ PIPETA na lékovku 20ml	17.36	14.20
549	KAPACÍ SKLENĚNÁ PIPETA na lékovku 50ml	18.18	13.36
550	KAPACÍ SKLENĚNÁ PIPETA na lékovku 100ml	19.01	13.55
556	KAPACÍ SKLENĚNÁ PIPETA na lékovku 30ml	17.36	15.05
672	DÁVKOVACÍ PUMPIČKA k výrobkům o objemu 500 ml	15.70	12.72
683	LÁHEV 0.5 l - plast	12.40	9.31
684	LÁHEV 1 l - plast	17.36	13.19

Filtrování textu

```
> sk[sk['varianta'].str.contains(" PET ")]        # vyhledání řetězce
                                      varianta  prodej  nakup
307            Láhev PET se sprejem, bílá, 100ml   23.97  20.63
311               Láhev PET se sprejem čirá 30ml   21.49  17.33
312               LÁHEV PET S UZÁVĚREM BÍLÁ 30ml   14.88  10.80
..                                          ...     ...    ...
315        LÁHEV PET S ODKLÁPĚCÍM UZ. ČIRÁ  200ml   23.14  17.58
447  LÁHEV PET S ODKLÁPĚCÍM UZ. ČIRÁ PLOCHÁ 100ml   13.22  11.00
580        LÁHEV PET S ODKLÁPĚCÍM UZ. bílá  200ml   23.14  17.89
> sk[sk['varianta'].str.contains(r"\bPET\b")]     # vyhledání regulárním výrazem
                                      varianta  prodej  nakup
307            Láhev PET se sprejem, bílá, 100ml   23.97  20.63
311               Láhev PET se sprejem čirá 30ml   21.49  17.33
312               LÁHEV PET S UZÁVĚREM BÍLÁ 30ml   14.88  10.80
..                                          ...     ...    ...
315        LÁHEV PET S ODKLÁPĚCÍM UZ. ČIRÁ  200ml   23.14  17.58
447  LÁHEV PET S ODKLÁPĚCÍM UZ. ČIRÁ PLOCHÁ 100ml   13.22  11.00
580        LÁHEV PET S ODKLÁPĚCÍM UZ. bílá  200ml   23.14  17.89
```

Ignorování velikosti písmen

```
> sk[sk['varianta'].str.contains(r"\bpet\b")]     # funkce rozlišuje velikost písmen
Empty DataFrame
Columns: [varianta, prodej, nakup]
Index: []

> sk[sk['varianta'].str.contains(r"\bpet\b",case=False)] # bez rozlišení velikosti
                                          varianta  prodej  nakup
307              Láhev PET se sprejem, bílá, 100ml   23.97  20.63
311                 Láhev PET se sprejem čirá 30ml   21.49  17.33
312                 LÁHEV PET S UZÁVĚREM BÍLÁ 30ml   14.88  10.80
313                 LÁHEV PET S UZÁVĚREM BÍLÁ 50ml   15.70  14.03
314                LÁHEV PET S UZÁVĚREM BÍLÁ 100ml   20.66  14.74
315         LÁHEV PET S ODKLÁPĚCÍM UZ. ČIRÁ  200ml   23.14  17.58
447  LÁHEV PET S ODKLÁPĚCÍM UZ. ČIRÁ PLOCHÁ 100ml   13.22  11.00
580         LÁHEV PET S ODKLÁPĚCÍM UZ. bílá  200ml   23.14  17.89
```

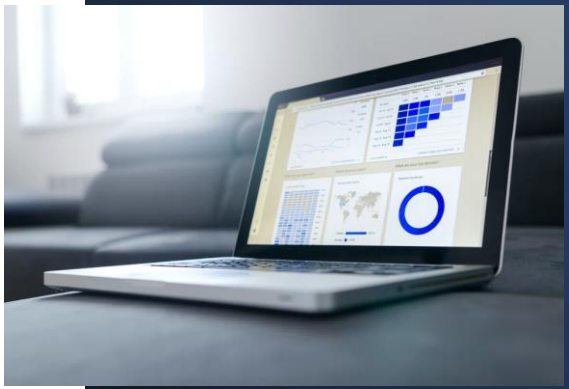
Kombinace filtru a výběru sloupců

```
> sk.loc[sk['varianta'].str.contains(r"\bpet\b",case=False),['varianta','prodej']]
```

	varianta	prodej
307	Láhev PET se sprejem, bílá, 100ml	23.97
311	Láhev PET se sprejem čirá 30ml	21.49
312	LÁHEV PET S UZÁVĚREM BÍLÁ 30ml	14.88
313	LÁHEV PET S UZÁVĚREM BÍLÁ 50ml	15.70
314	LÁHEV PET S UZÁVĚREM BÍLÁ 100ml	20.66
315	LÁHEV PET S ODKLÁPĚCÍM UZ. ČIRÁ 200ml	23.14
447	LÁHEV PET S ODKLÁPĚCÍM UZ. ČIRÁ PLOCHÁ 100ml	13.22
580	LÁHEV PET S ODKLÁPĚCÍM UZ. bílá 200ml	23.14

PANDAS
svatý grál analýzy dat

Agregace dat



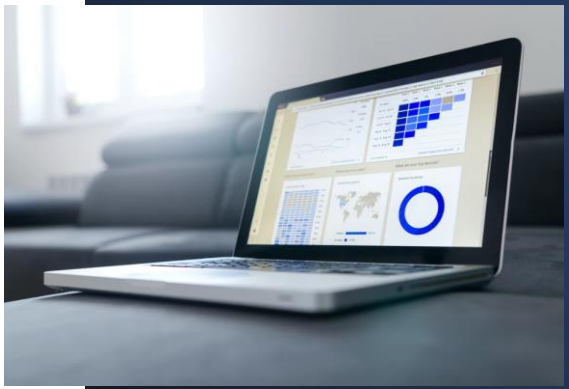
Agregace dat

```
> on_stock = sklad.groupby('produkt')['skladem'].sum()
> on_stock.sort_values()
produkt
extrakt        3
katalog        4
kabelka        9
cat           14
garmin        15
              ...
balzám       893
samsung     1099
apple       1755
olej        6892
esence     19684
Name: skladem, Length: 70, dtype: int64
```

PANDAS
svatý grál analýzy dat

Přidávání sloupců

Spojování tabulek



Spojení tabulek „UNION“

```
> top_sellers = sklad.groupby('produkt')['prodano'].sum().sort_values(ascending=False).head(5)
> low_sellers = sklad.groupby('produkt')['prodano'].sum().sort_values(ascending=False).tail(5)

> toplow_sellers = pandas.concat([top_sellers,low_sellers],axis=0)
> toplow_sellers
produkt
esence       23364
olej          8048
apple         1863
samsung       1287
lékovka        931
nová            30
katalog         30
comtrend        21
testovací       15
kabelka         11
Name: prodano, dtype: int64
```

Spojení tabulek
podle společné hodnoty

```
> sold = sklad.groupby('produkt')['prodano'].sum()
> stocksold = pandas.merge(on_stock,sold,how="left",on="produkt")
> stocksold
```

produkt	skladem	prodano
alcatel	193	138
alfalfa	152	327
apple	1755	1863
aroma	387	419
aroma-andílek	89	63
...	...	...
čichová	101	42
ředkev	462	536
ředící	130	267
řepa	630	792
řeřicha,	275	312

Přidání nového sloupce

```
> sklad['hodnota'] = sklad['skladem'] * sklad['nakup']
> sklad['vynos'] = sklad['prodano'] * sklad['prodej']
> sklad['zisk'] = sklad['prodej'] - sklad['nakup']
> sklad.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 807 entries, 0 to 806
Data columns (total 12 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   skupina   807 non-null    object
 1   pid       807 non-null    object
 2   produkt   807 non-null    object
 3   varianta  807 non-null    object
 4   prodej    807 non-null    float64
 5   nakup     807 non-null    float64
 6   skladem   807 non-null    int64
 7   prodano   807 non-null    int64
 8   marze     807 non-null    float64
 9   hodnota   807 non-null    float64
 10  vynos     807 non-null    float64
 11  zisk      807 non-null    float64
dtypes: float64(6), int64(2), object(4)
memory usage: 82.1+ KB
```

Přidání sloupců voláním funkce

```
def build_new_columns(key,row):          # definice funkce
  ...
  output = [key,id,produkt,mnozstvi]     # výstupem je list hodnot,
  return output                          # které se uloží do DF


df[['skupina','id','produkt','množství']] = df.apply(
              lambda row: build_new_columns('kosmetika',row),
              axis=1,                    # pracujeme s řádky
              result_type='expand‘       # rozdělí list do sloupců
            )
```