pandas - dátové rámce

Načítanie dát

data = pd.read_csv("subor.csv") – načítanie dát z CSV súboru
data = pd.read_excel("subor.xlsx") – načítanie dát z Excel súboru

Veľkosť tabuľky

`len(data)` `data.shape[0]` – počet riadkov
`data.shape[1]` – počet stĺpcov

Výber stĺpcov

`data.columns` – typ a názvy stĺpcov
`stlpec = data["stlpec"]` – výber jedného stĺpca
`data = data[["stlpec1", "stlpec2"]]` – výber viacerých stĺpcov
`data = data.drop(columns=["stlpec1", "stlpec2"])` – odstránenie stĺpcov

Výber riadkov

```
vyber = data.query("Krajina == Slovensko")
```

dopyt je reťazec s testom porovnávajúcim hodnoty atribútov, okrem == je možné použiť !=, <, >, <=, >=, in, zložitejšie výrazy je možné vyjadriť spojkami not, and, or.

Spájanie dát po riadkoch

```
data = pd.concat([data_1, data_2], ignore_index=True, sort=False)
```

Spájanie dát po stĺpcoch podľa kľúča

```
data_vsetky = pd.merge(left=data_2015, right=data_2016,
left_on="Krajina", right_on="Krajina", how="left")
```

Metódy spájania left, right, outher, inner – iba z ľavej, iba z pravej, zjednotenie oboch, prienik oboch

Základné štatisticky a výpis dát

`data.head()` `data.tail()` – prvé/posledné riadky z tabuľky (štandardne 5 riadkov)
`data.describe()` – súhrné štatistiky pre celú tabuľku
`data["stlpec"].describe()` – súhrné štatistiky pre jeden stĺpec
`min()` `max()` `count()` `mean()` `std()` `quantile(n)` – samostatné štatistiky, minimum, maximum, počet nechýbajúcich hodnôt, stredná hodnota, štandardná odchýlka, kvartil (n=0.25, 0.5, 0.75)
`data["stlpec"].value_counts()` – početnosť hodnôt pre diskrétny atribút

Spracovanie chýbajúcich hodnôt

`data["stlpec"].isna().sum()` – zistenie počtu chýbajúcich hodnôt pre atribút
`data["stlpec"] = data["stlpec"].fillna(hodnota)` – nahradenie chýbajúcich hodnôt konštantou (napr. priemerom)

Výpočet/zmena hodnôt

```
data_2016["Score"] = data_2016.eval("GDP + Family + Health + Freedom")
```

Výpočet hodnôt podľa zadaného výrazu

```
data["vysledok"] = data["stlpec"].apply(funkcia)
```

Aplikovanie funkcie na hodnotu stĺpca a uloženie výsledných hodnôt do stĺpca

```
vysledok = data.apply(funkcia, axis=1)
```

Aplikovanie funkcie po riadkoch a uloženie výsledných hodnôt do stĺpca

Kontingenčné tabuľky

```
tabulka = pd.pivot_table(data, index="Region", columns="Rok",
values="Pocet")
```

index určuje riadky tabuľky, columns určuje stĺpce a values polia, pre každý argument môže byť uvedených viacero stĺpcov, štandardná agregačná funkcia je priemer

```
tabulka = pd.pivot_table(data, index="Region", values=["Pocet",
"Umiestenie"], aggfunc={"Pocet": "mean", "Umiestnenie": ["min", "max"]})
```

Tabuľka s viacerými agregačnými funkciami

```
tabulka = pd.crosstab(index=[data["Mesto"], data["Vzdelanie"]],
columns=data["Pohlavie"])
```

pre kategorické atribúty vypočíta tabuľku s počtami rôznych kombinácií hodnôt

Korelačná analýza

`tabulka = data.corr()` – vypočíta korelačnú maticu pre všetky číselné atribúty