Slovenská technická univerzita v Bratislave

Fakulta informatiky a informačných technológií

Ilkovičova 2, 842 16, Bratislava 4

Vizualizácia dát
Prvá úloha: Vizualizácia teploty
Python: Bokeh

Autor: Bc. Peter Berta

Cvičenie: Ing. Patrik Polatsek (štvrtok o 19:00)

Semester: Letný semester 2018

1. Cieľ:

Otestujte zobrazenie zmeny teploty vo vybranej oblasti pomocou grafu. Vyhodnoťte efektivitu tejto vizualizačnej metódy na malých, stredných a veľkých dátach. Vyvoďte záver z pozorovaných výsledkov.

2. Dáta:

Dataset:

Ako testovacie dáta som si zvolil dataset[1], v ktorom je zaznamenaná priemerná denná teplota pre rôzne mestá sveta. Niektoré mestá majú dostupné dáta až do roku 1743. Formát súboru je *.csv*.

Vzhľadom na obmedzený počet záznamov z jedného mesta som sa rozhodol zobraziť aj viacero miest v jednom grafe. Pre vizualizáciu som si zvolil nasledovné mestá:

- Odessa - Ukraine
- Paris - France
- Stockholm - Sweden
- Uppsala – Sweden

Chýbajúce dáta:

Niektoré dni nemajú zaznamenanú teplotu. Chýbajúce dáta sú doplnené hodnotami v ich okolí.

```
# fill missing data
df3 = df2.fillna(method='bfill', limit=50).fillna(method='ffill', limit=50)
```

[1] Zdroj dát: https://figshare.com/articles/temperature_csv/3171766

3. Implementácia:

Pre implementáciu tejto úlohy som si zvolil programovací jazyk Python[2], knižnicu Bokeh[3].

- Inicializácia dát: pomocou knižnice Pandas[4] som prečítal dáta o teplotách z **.csv* súboru
- Filtrovanie dát: z celého datasetu som si vyfiltroval len podstatné informácie

```
# filter important columns
df2 = df.filter(items=['month', 'day', 'year',
                       'AverageTemperatureFahr', 'City'])
```

- Vyplnenie chýbajúcich dát: popísané vyššie
- Konverzia dátumu na formát *datetime:* tento krok je obzvlášť dôležitý, aby bolo možné pridať dátum na jednu z osí

```
df3 = df3.assign(Date=pandas.Series(
    pandas.to_datetime({'year': df3.year,
                        'month': df3.month,
                        'day': df3.day})).values)
```

- Zobrazenie dát: podľa voľby sa zobrazia dáta rôzneho rozsahu (popísané vo výsledkoch)

```
plot = figure(plot_width=500, plot_height=400,
              x_axis_type="datetime",
              title='Temperature (F)')
..
plot.legend.location = "bottom_left"
plot.border_fill_color = "whitesmoke"
output_file('small.html')
```

- Zobrazenie jednej čiary: tento krok je silno viazaný k predchádzajúcemu kroku. V každom zobrazení je ale tento krok mierne upravený, rozšírený o ďalšie čiary.

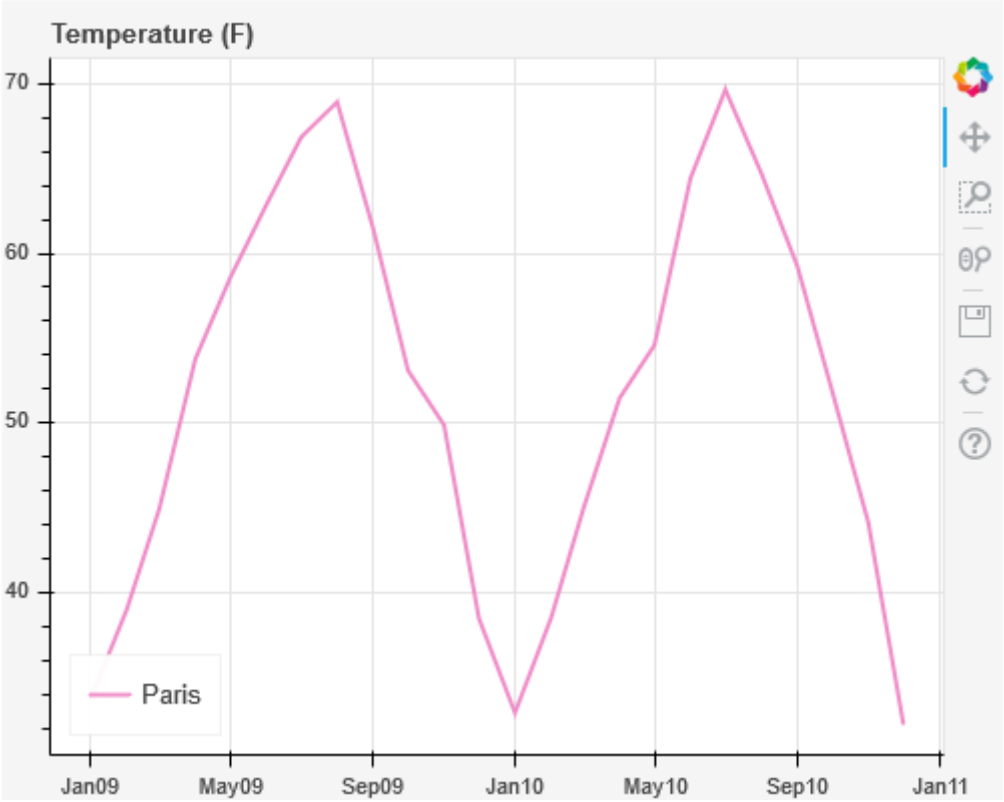
```
plot.line(df4[(df4.City == 'Odesa') &
              (df4.Date >= start_date) &
              (df4.Date < end_date)].Date,
          df4[(df4.City == 'Odesa') &
              (df4.Date >= start_date) &
              (df4.Date < end_date)].AverageTemperatureFahr,
          legend='Odesa',
          line_color='#5bb2ff',
          line_width=2,
          line_cap='round')
```

[2] Web: https://www.python.org/
[3] Web: https://bokeh.pydata.org/en/latest/
[4] Web: https://pandas.pydata.org/

4. Výsledky:

Malé dáta (730 záznamov):

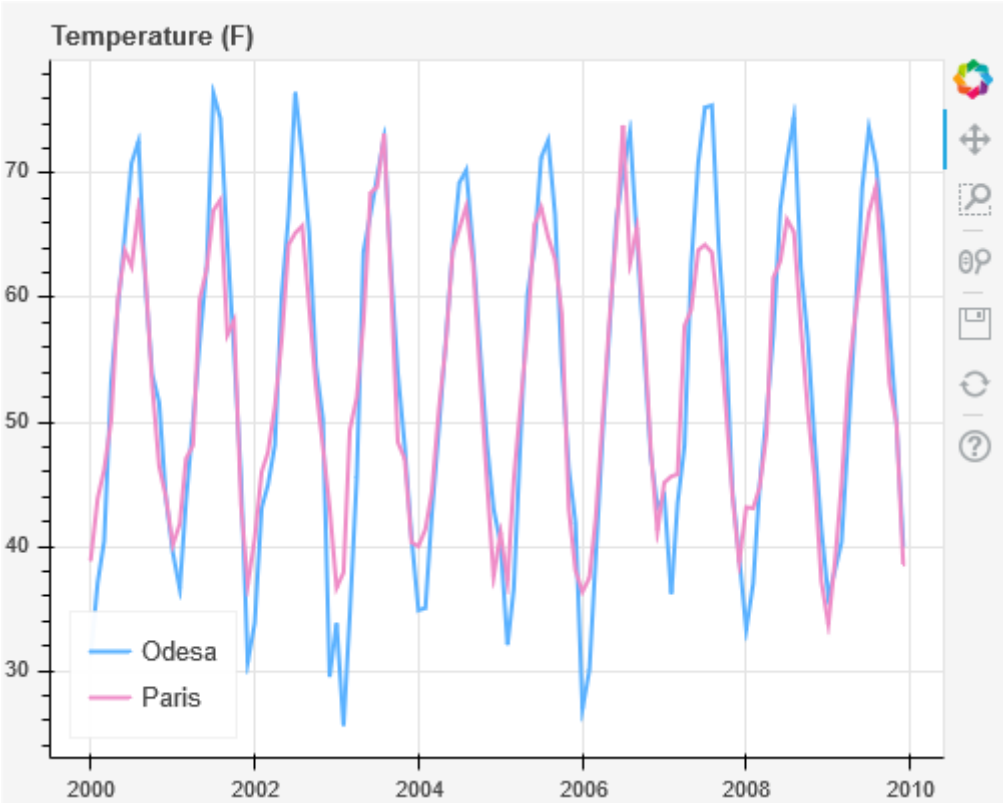
Ako malý dataset som si zvolil priemerné teploty v Paríži za roky 2009 a 2010. Tieto dáta považujem za najprehľadnejšie, i keď nie veľmi zaujímavé. Je možné pozorovanie zmeny teploty počas roka, alebo porovnanie teploty dvoch rokov.



Obr.1: Teplota v Paríži (2009-2011)

Stredne veľké dáta (7300 záznamov):

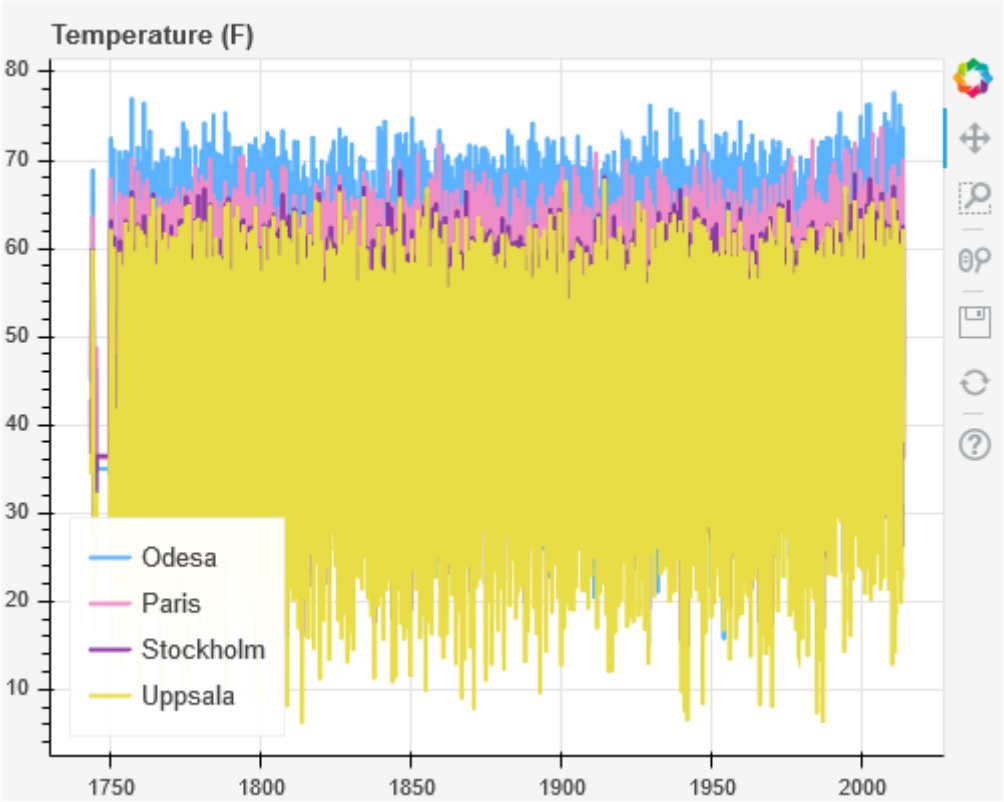
Ako stredne veľké dáta som si zvolil mestá Paríž a Odesu, teraz ale za obdobie od roku 2000 do roku 2010 (Obr.1). Táto vizualizácia je podľa mňa oveľa zaujímavejšia, vzhľadom na to, že je možné porovnanie dvoch rôznych miest, a taktiež aj rokov za sebou. Jednotlivé záznamy od seba nie je možné rozlíšiť, za to je ale možné sledovať ich celkový priebeh. V prípade potreby je vždy možné si jednotlivé časti grafu priblížiť pomocou funkcie *zoom*.



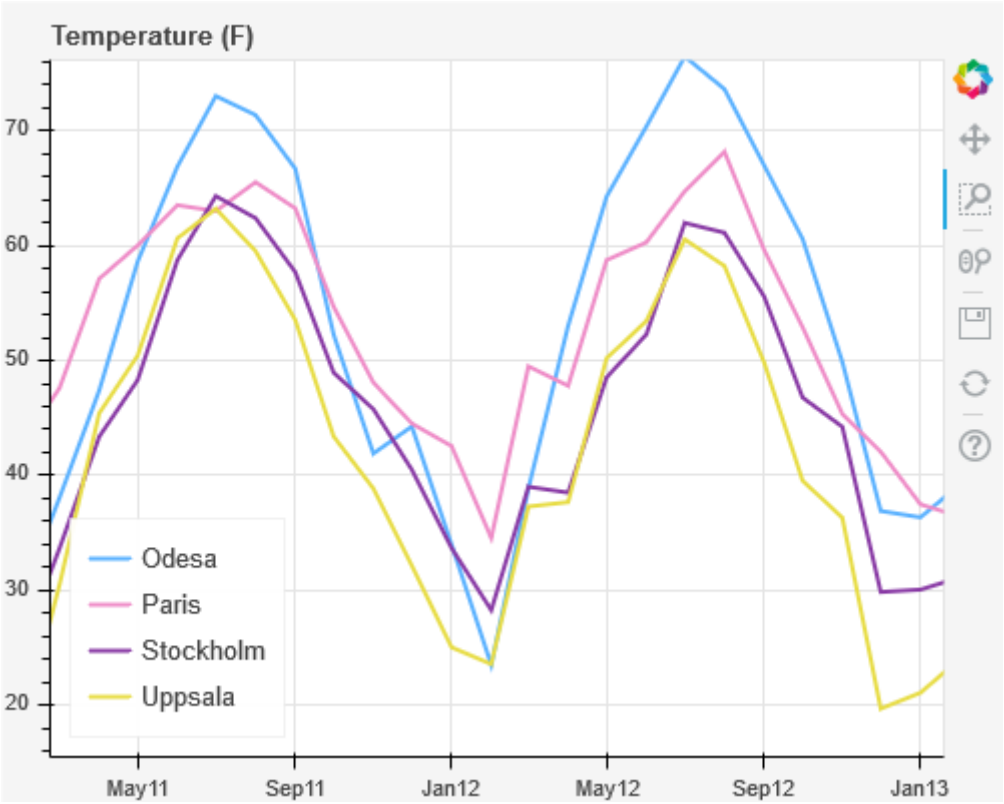
Obr.2: Teplota v Paríži a v Odese (2000-2010)

Veľké dáta (13000 záznamov):

Pre zobrazenie veľký dát som si zvolil informácie o teplote v mestách Odesa, Paríž, Štokholm a Uppsala (Obr. 3). Teraz už som dáta časovo neobmedzoval. Z toho dôvodu dáta začínajú v roku 1743. Teraz už je graf vysoko neprehľadný a nie je možné vyvodiť žiadne závery. Použitím nástroja *zoom* je však možné obmedziť pozorované časové obdobie na menšiu časovú jednotku. V takom prípade je možné porovnať teplotu v štyroch mestách napríklad počas roka. Takýto náhľad je zobrazený na obr.4.



Obr.3: Obr.2: Teplota v Odese, Paríži, Štokholme a v Uppsale (1743-2013)



Obr.4: Ilustračný zoom na veľké dáta

5. Zhodnotenie:

Grafové zobrazenie určite má svoje miesto v oblasti vizualizácii dát. Ako sme ale mohli vidieť, je dôležité vhodne zvoliť rozsah zobrazovaných dát. Taktiež je potrebné dbať na farby jednotlivých čiar. Ak by boli čiary, ktoré spájajú jednotlivé teplotné záznamy príliš podobnej farby, mohli by sa pliesť a pozorovateľa odradiť.

Tento projekt bol pre mňa zaujímavý aj vzhľadom na to, že som sa zoznámil s knižnicou *Bokeh* a taktiež som sa zoznámil s využitím *Jupyter Notebooku*[5].

[5] Web: http://jupyter.org/