Czech Technical University in Prague

Faculty of Electrotechnical Engineering

Department of Cybernetics
Department of Computer Science

Vytěžování dat – cvičení II

Načtení dat, vizualizace

Jan Drchal, Oleg Kovářík, Petr Pošík, Pavel Kordík

Osnova cvičení

- Import dat
  - Matice
  - Dataset (statistický toolbox)
- Normalizace dat
- Histogram
- Bodové grafy

Načtení dat - matice

- UCI repository - databáze automobilů – moodle

```
auta_mat = csvread('auto-numbers-only.csv');
```

- Načte data jako matici
  - data nesmí obsahovat chybějící hodnoty a nominální hodnoty
  - Všechny atributy numerické
  - Soubor neobsahuje jména atributů

Normalizace dat – v1 (matice)

Matice po sloupcích – vytvořte následující soubory:

- Obsah souboru norm01v1.m

```
function data = norm01v1(data)
    for i = 1:size(data,2),
        data(:,i) = minmax(data(:,i));
    end
end
```

- Obsah souboru minmax.m

```
function vec = minmax(vec)
    vec = (vec - min(vec)) / (max(vec) - min(vec));
end
```

Úkol: použijte funkci norm01v1 na auta_mat

Načtení dat - dataset

- stáhnout dataset auto-mpg.data-mod
- objekt dataset ze Statistics Toolboxu

```
auta_dat = dataset('file','auto-mpg.data-mod-names.csv', ...
                   'ReadVarNames', false, ...
                   'ReadObsNames', false, ...
                   'delimiter', ',');
```

- Načte data ve formě datasetu
  - Může obsahovat jména atributů (`'ReadVarNames'`)
  - Může obsahovat chybějící hodnoty (`'TreatAsEmpty'`)
  - Může obsahovat nominální data
  - Atributy lze dodatečně pojmenovat

```
auta_dat(1:5, :)
```

Atributy a pojmenování dat

Atributy:

- mpg: miles-per-galon, počet mil ujetých na 1 galon paliva
- cyl: cylinders, počet válců
- disp: displacement, zdiv
- hp: horsepower, koňských sil
- wgt: weight, hmotnost
- acc: acceleration, zrychlení
- year: rok výroby
- org: origin, původ (1 - Amerika, 2 - Evropa, 3 - Japonsko)

- pojmenování atributů datasetu

```
auta_dat = set(auta_dat, ...
  'VarNames', {'mpg', 'cyl', 'disp', 'hp', ...
  'wgt', 'acc', 'year', 'org', 'name'});
auta_dat(1:5, :)
```

Převody, souhrn

- Převedení proměnné org na nominální (1 - Amerika, 2 - Evropa, 3 - Japonsko)

```
tmporg = nominal(auta_dat.org,
   {'America','Europe','Japan'});
auta_dat = replacedata(auta_dat, tmporg, 'org');
auta(1:5,:)
```

- Převod dat z datasetu do numerické matice (jen u numerických, ordinálních a nominálních proměnných; zde např. nelze převést proměnnou *name*)

```
double(auta_dat(:,1:8))
```

- Počáteční průzkum dat

```
summary(auta_dat)
```

Normalizace dat – v2 (dataset)

- funkce minmax pro jeden vektor (soubor minmax.m)

```
function x01 = minmax(x)
x01 = (x - min(x)) / (max(x) - min(x));
end
```

- aplikace na sloupce 1-7 datové sady

```
auta01 = auta_dat;
x01 = datasetfun( @minmax, auta01(:,1:7), ...
                  'UniformOutput', false );
x01 = [x01{:}]; % Převod cell array na matici
auta01 = replacedata( auta01, x01, 1:7);
mins = datasetfun( @min, auta01(:,1:7) )
maxs = datasetfun( @max, auta01(:,1:7) )
```

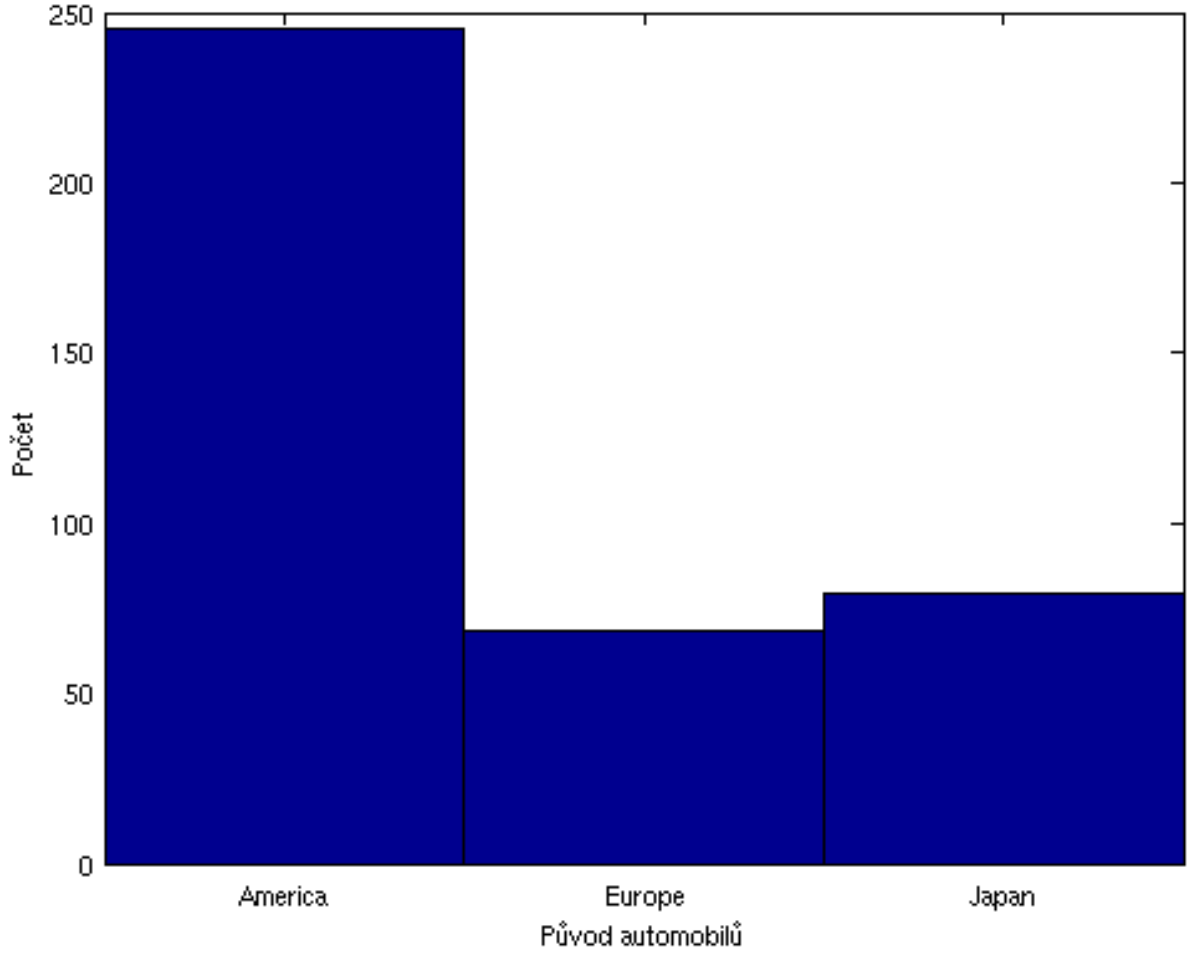
Histogram

- V případě maticových dat zaměňte `auta_mat(:,8)` za `auta_dat.org`
- Četnost bodů v kategoriích.

```
kat = getlabels( auta_dat.org );
stredy = 1:numel(kat);
hist( double(auta_dat.org), stredy );
set(gca, 'XTickLabel', kat);
xlabel('Původ automobilů');
ylabel('Počet');

  %zkuste File/Publish v Editoru
```

Histogram II



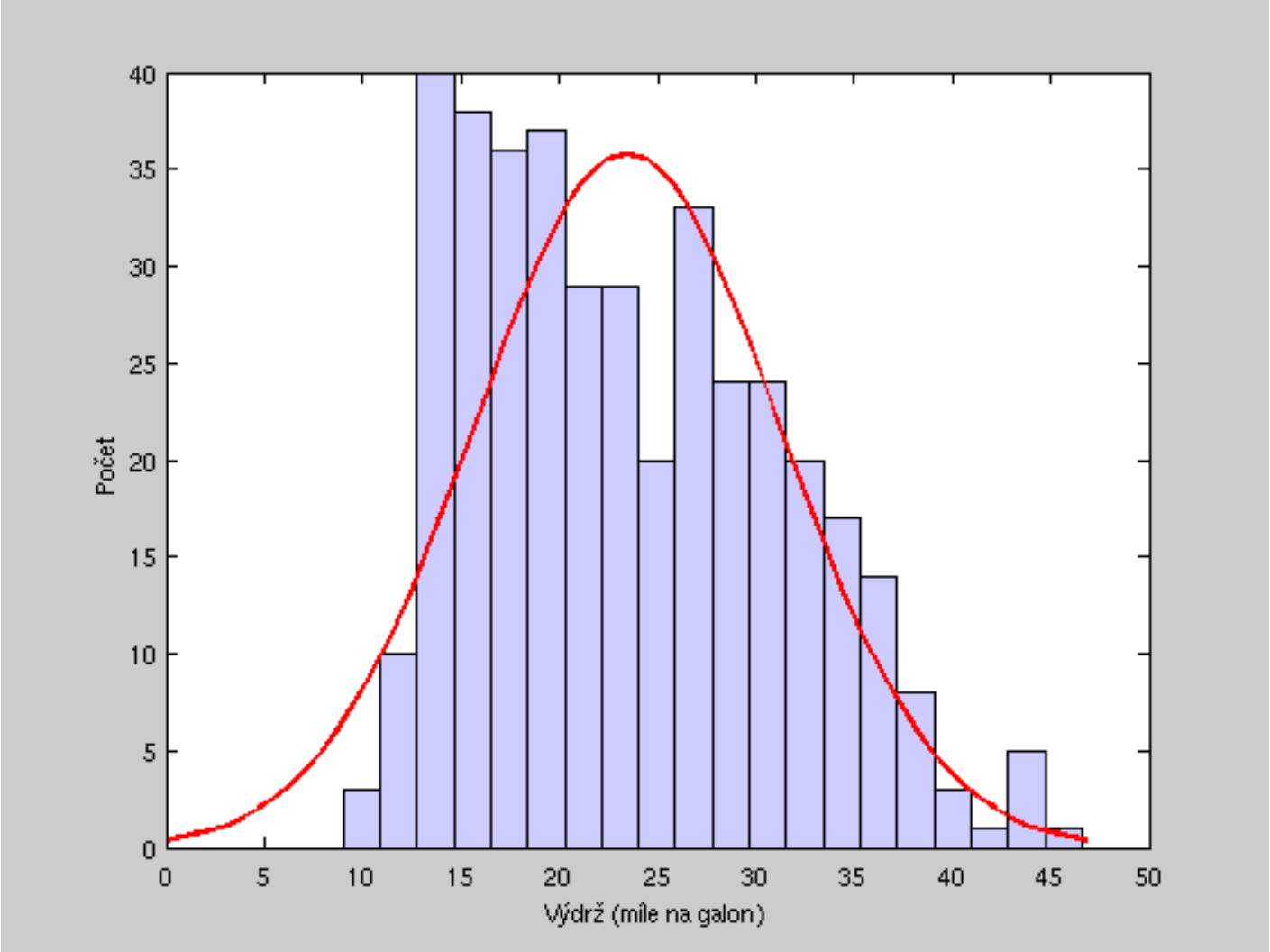
- Co z obrázku můžeme říct o datech?

Histogram pro spojité proměnné

- `histfit` – histogram s křivkou normálního rozdělení:

```
h = histfit(auta_dat.mpg);
xlabel('Výdrž (míle na galon)');
ylabel('Počet');
set(h(1),'FaceColor',[.8 .8 1]);
```

Histogram pro spojité proměnné II



Příklad:

- Udělejte 3 histogramy pod sebe – výdrž pro americké, evropské a asijské vozy zvlášť

- hint: americké vozy vyberete příkazem

```
auta_dat(auta_dat.org == 'Amerika', :)
```

- vytvoří se dataset obsahující
  - všechny proměnné (:)
  - jen ty řádky, pro něž je v proměnné *org* hodnota 'Amerika'

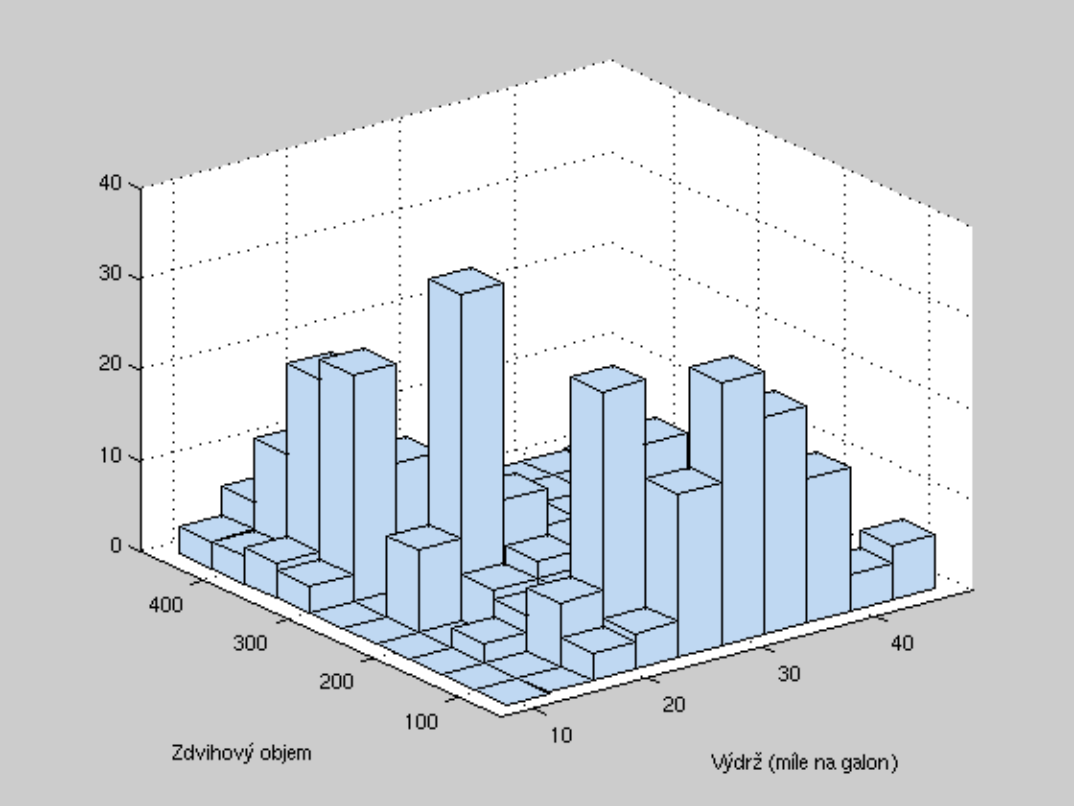
Příklad: Řešení

- Udělejte 3 histogramy pod sebe – výdrž pro americké, evropské a asijské vozy zvlášť

```
kat = getlabels( auta_dat.org );
for i = 1:numel(kat),
    subplot(numel(kat), 1, i);
    hist( auta_dat.mpg(auta_dat.org == kat{i}), 37 );
    axis([9 47 0 23]); title(kat{i});
end
```

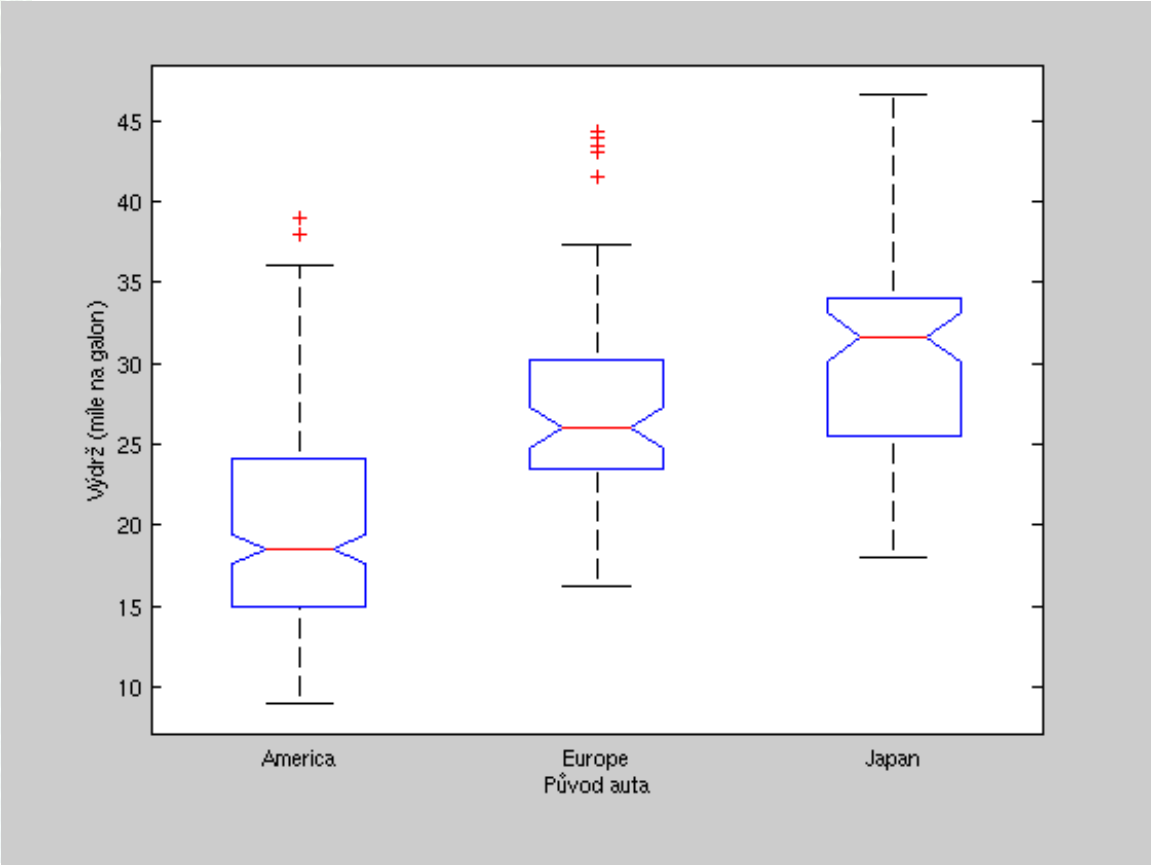
2D histogram

```
hist3([auta_dat.mpg, auta_dat.disp]);
xlabel('Výdrž (míle na galon)');
ylabel('Zdvihový objem');
```



Krabicový graf (boxplot)

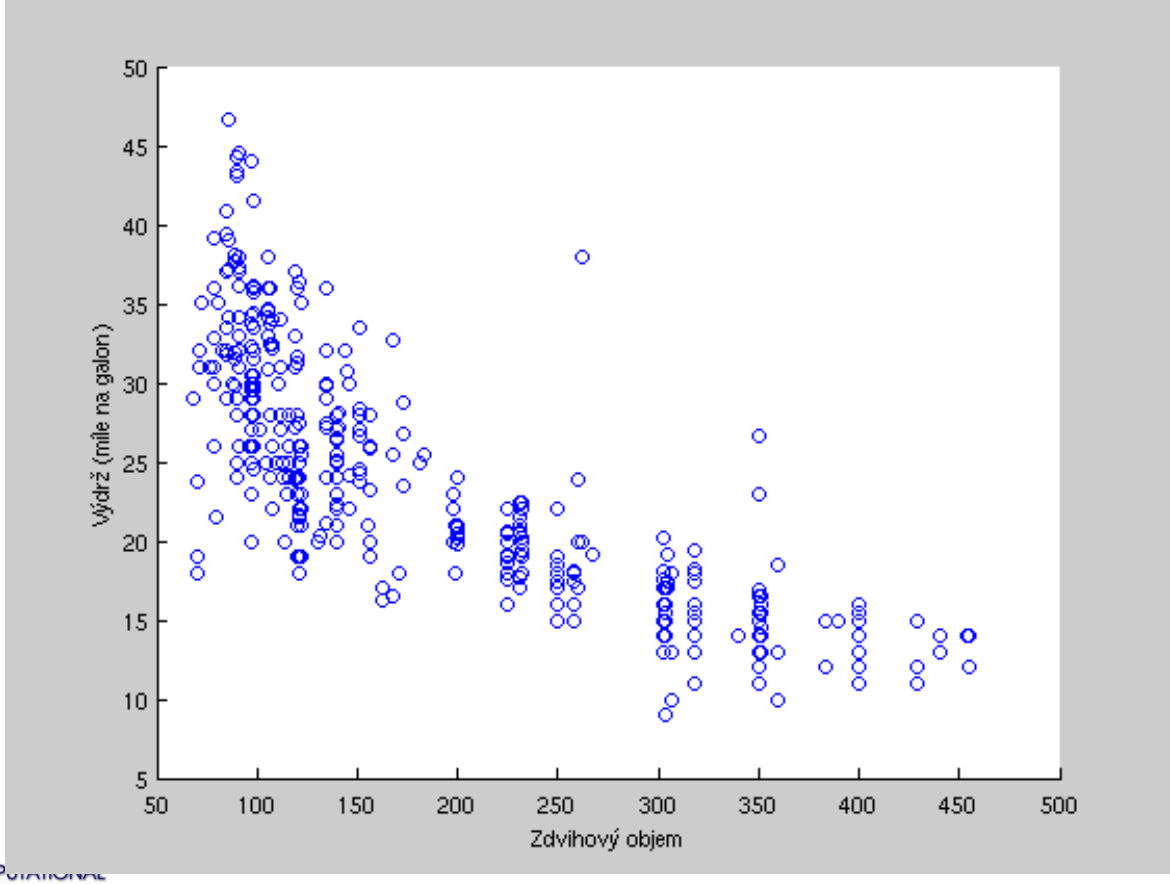
```
boxplot(auta_dat.mpg, auta_dat.org, 'notch', 'on');
xlabel('Původ auta');
ylabel('Výdrž (míle na galon)');
```



- červená čára – medián,
- pokud se zářezy nepřekrývají, je rozdíl mediánů mezi skupinami statisticky významný.

Bodový graf (scatterplot)

```
scatter(auta_dat.disp, auta_dat.mpg);
xlabel('Zdvihový objem');
ylabel('Výdrž (míle na galon)');
```

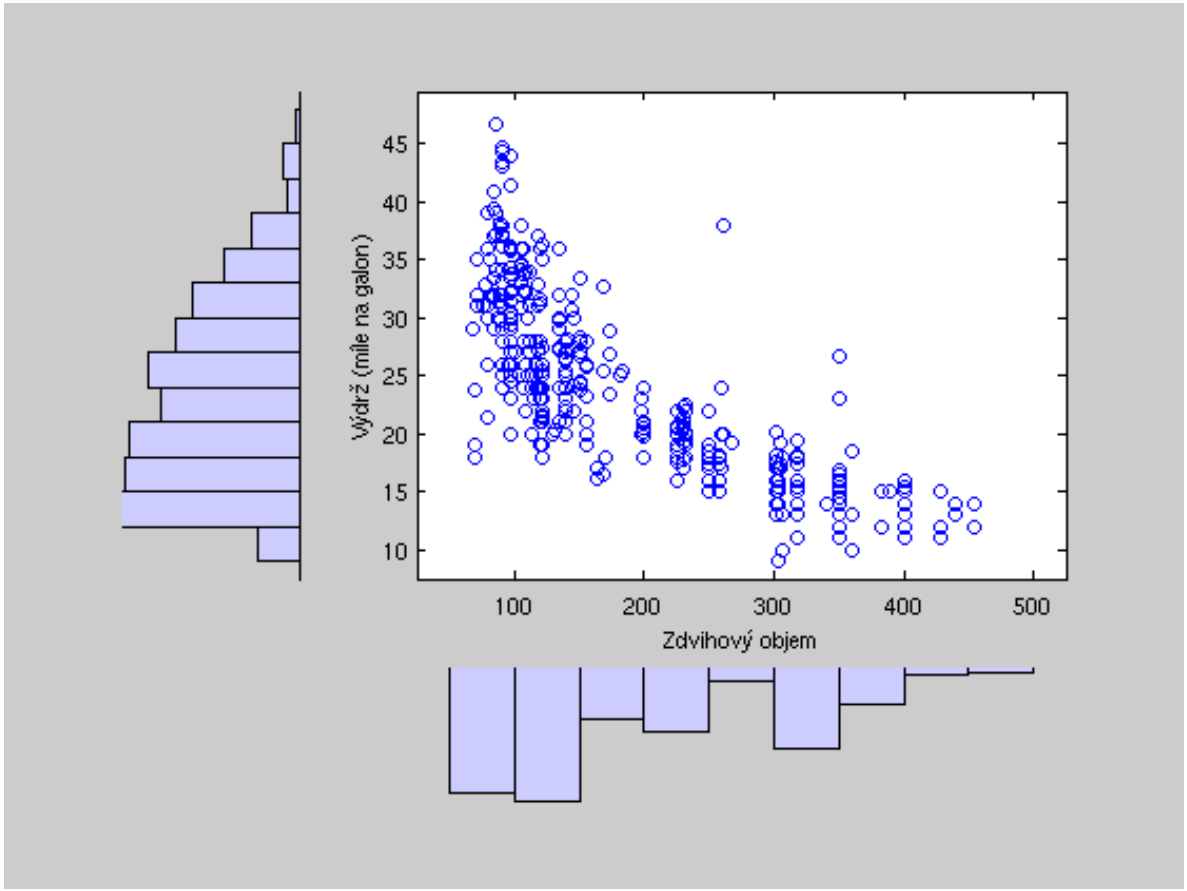


Můžeme použít i `plot`.

Co z grafu vyplývá?

Bodový graf s histogramy

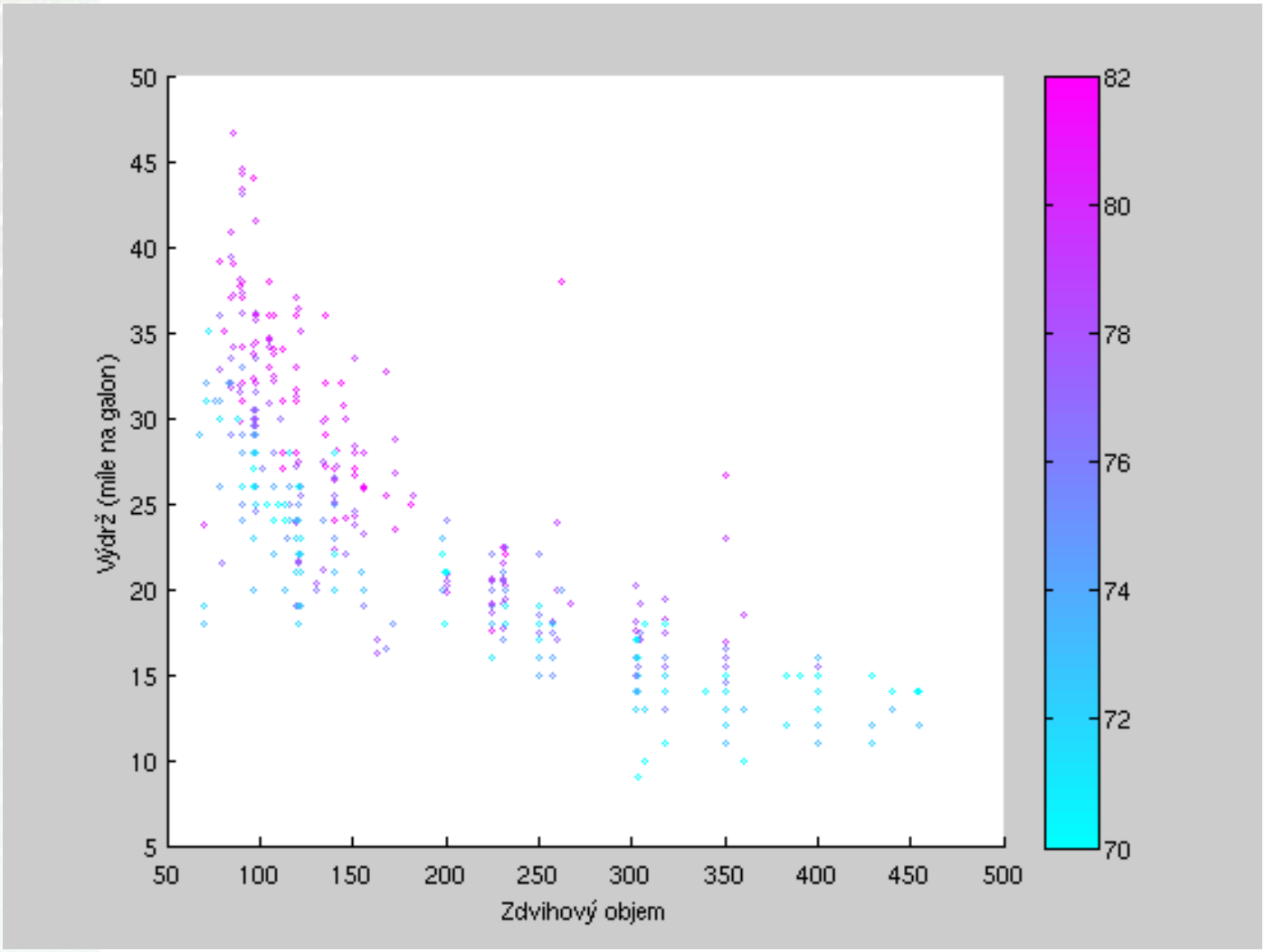
```
scatterhist(auta_dat.disp, auta_dat.mpg );
xlabel('Zdvihový objem');
ylabel('Výdrž (míle na galon)');
```



Bodový graf s barevným kódováním

```
scatter(auta_dat.disp, auta_dat.mpg,
  8, auta_dat.year);
colorbar;
colormap cool;
xlabel('Zdvihový objem');
ylabel('Výdrž (míle na galon)');
```

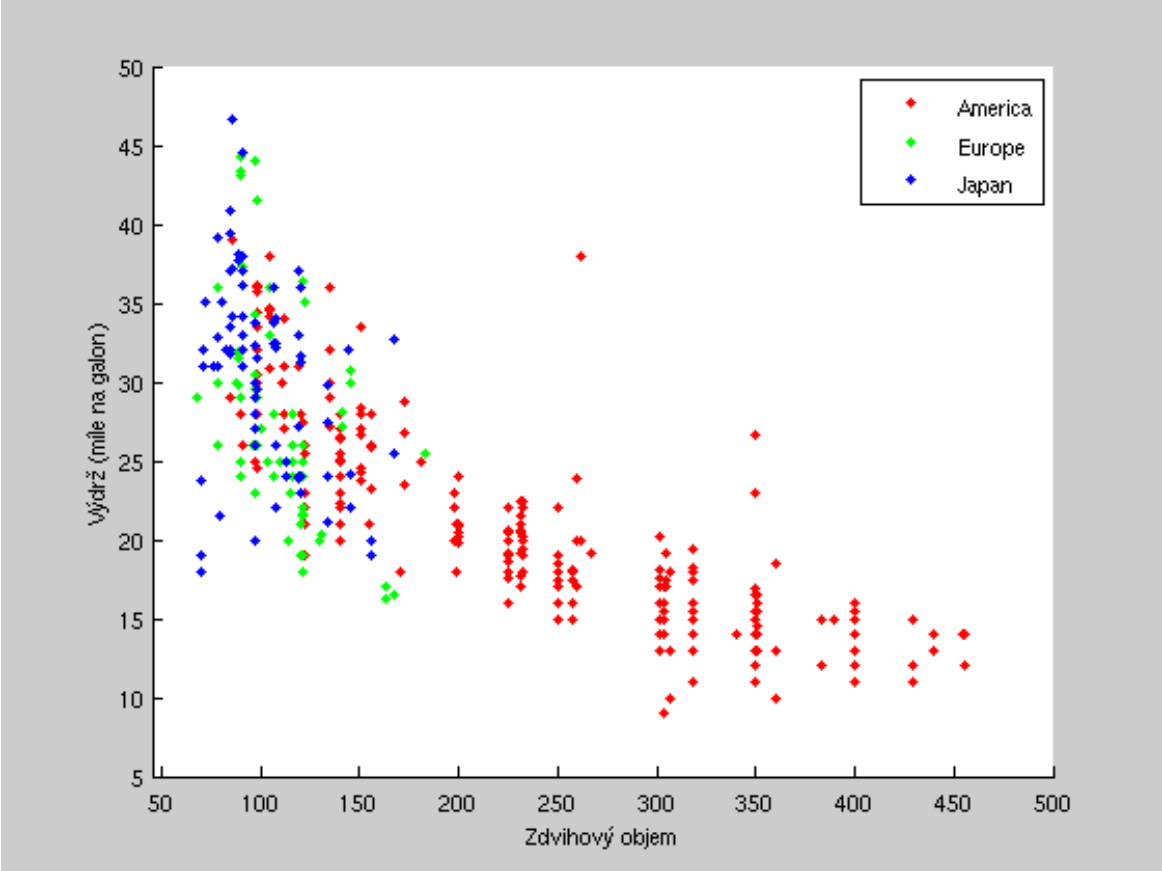
Bodový graf s barevným kódováním II



Co můžeme říct o vývoji zdvihového objemu a výdrže v 70. a 80. letech?

Kategorizovaný bodový graf

```
gscatter( auta_dat.disp, auta_dat.mpg,... auta_dat.org );
xlabel('Zdvihový objem');
ylabel('Výdrž (míle na galon)');
```



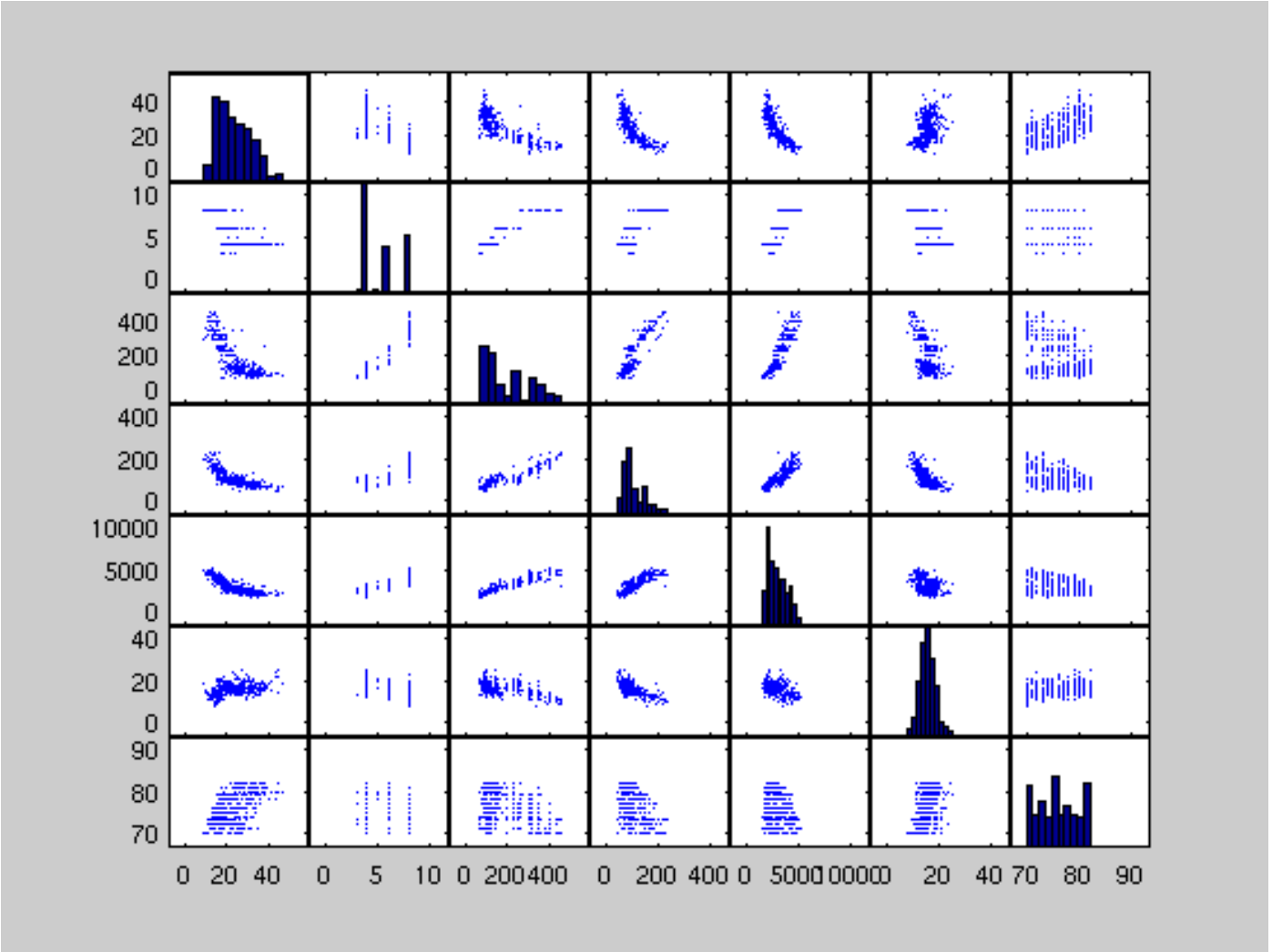
Matice bodových grafů

- Jak zobrazit mnohodimezionální data?
- `plotmatrix` (auta_mat)
- `gplotmatrix` (auta_dat)

```
auta_mat = double( auta_dat(:,1:7) );
plotmatrix(auta_mat);

promenne = get(auta_dat,'VarNames');
promenne = promenne(1:7);
```

Matice bodových grafů II



Matice bodových grafů III

```
gplotmatrix( auta_mat(:,1:4), auta_mat(:,5:7), auta_dat.org,
   [], [], 5, [], [], promenne(1:4), promenne(5:7) );
```