**Cvičenie 5  
Klasterizácia**

**Zadanie:**

Vyskúšajte na klasterizáciu využiť len časti z prideleného datasetu (20, 50, 75 percent...). Zmení sa nejako ideálny počet klastrov? Ako potom vyzerá graf s klasterizovanými údajmi?

**Dataset:**

Poskytnutý dataset „Mall\_Customers.csv“ obsahuje ID zákazníka, pohlavie, vek, ročný príjem a skóre hodnotiace výdaje na nákup.

**Riešenie:**

Pri bližšom skúmaní vidíme, že dataset je usporiadaný nielen podľa ID zákazníka ale aj jeho ročného príjmu. Preto pri skúšaní klasteriácie využitím len časti prideleného datasetu sme postupovali najskôr tak, že sme brali len prvých 20, 50, 75 percent. No po uvedomení si, že dáta sú usporiadané podľa príjmov sme najskôr dataset premiešali a až následne vyberali potrebné percento dát. Pre zaujímavosť uvádzame aj prvý pokus, kde sme dáta nepremiešali.

A graph of a number of clusters

Description automatically generatedA screen shot of a graph

Description automatically generatedPokus 1:

**A screen shot of a graph

Description automatically generatedA graph of a number of clusters

Description automatically generated**

**A screen shot of a graph

Description automatically generatedA graph of a number of clusters

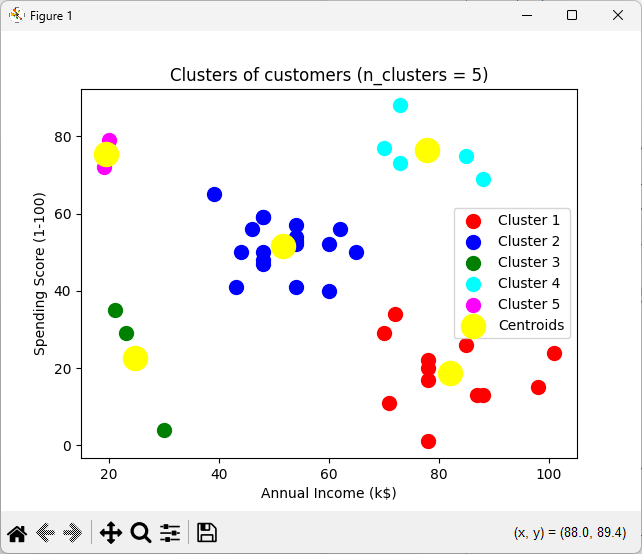
Description automatically generated**

Pokus 2:

A graph on a screen

Description automatically generatedA graph on a computer screen

Description automatically generatedA graph on a computer screen

Description automatically generated

A screen shot of a computer screen

Description automatically generated

A screen shot of a graph

Description automatically generated

**Zhrnutie:**

Ako môžeme vidieť počet klastrov sa pri pokuse číslo 2 výrazne nezmenil a to z dôvodu zastúpenia každej skupiny zákazníkov, keďže sme dataset najskôr premiešali a až následne vyberali určité percento ľudí. Pri pokuse číslo 1 sa počet klastrov líšil, nakoľko sme vyberali len prvých niekoľko percent zo zoradených údajov a preto bol počet klastrov menší a v prípade 75% vzorky rovnaký. V niektorých prípadoch môže byť diskutabilné zvolenie počtu klastrov, napríklad pri pokuse dva pre 50% by možno niekto uviedol ideálny počet 4, my sme však zvolili 6 čo nám oddelí ďalšiu skupinu zákazníkov.