



Strojové učenie a neurónové siete - **zadanie 4** - učenie bez učiteľa (zhlukovanie)

Zadanie 4

- Zadanie a bodovanie je k dispozícii na dokumentovom serveri v AIS
- Máte sa naučiť:
 1. Exploratory Data Analysis
 2. Pracovať s textovými dátami
 3. Zhlukovať dáta
 4. Zmenšovať dimenziu dát
- Dataset:
 - Hry na Steame, dostupný v AIS.



Obsah tejto prezentácie

- EDA
 - Učenie bez učiteľa
 - Typy zhľukovania
 - Vybrané zhľukovacie algoritmy - kMeans, DBSCAN, AHC, GMM
-
- Ukážka - sklearn, plotly, pandas,
 - O týždeň: redukovanie dimenzie



Učenie bez učiteľa

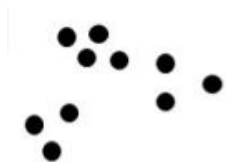
- Hľadá vzory v dátach
- Definujeme vstupnú množinu a parametre modelu
- Ako výstup získavame napríklad:
 - Príslušnosti ku zhlukom (“prirodzené” triedy)
 - Lepšie pochopenie dát/vzťahov
 - Iná reprezentácia dát
- Využitie učenia bez učiteľa:
 - Nedostatok označených dát na tréningovanie
 - Data mining - nevieme koľko/aké dáta (triedy) máme
 - Pochopenie dát - napr. Exploratory Data Analysis
 - “Feature engineering”
- Oproti učeniu s učiteľom:
 - + Ľahšie získať dáta
 - Obvykle výpočtovo komplexnejšie
 - Ako vyhodnocujem úspešnosť?



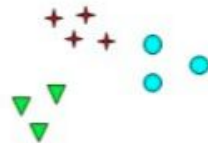
- Príklady:
 - Zhlukovacie algoritmy
 - Anomálie
 - Autoenkódery



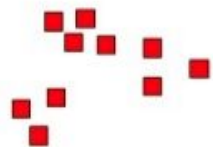
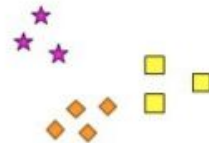
Učenie bez učiteľa



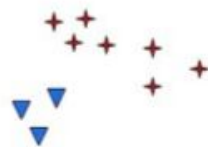
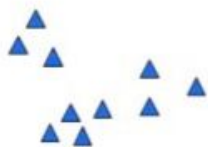
How many clusters?



Six Clusters



Two Clusters

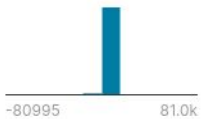
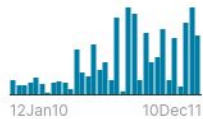

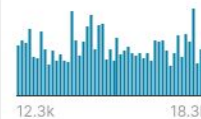



Four Clusters



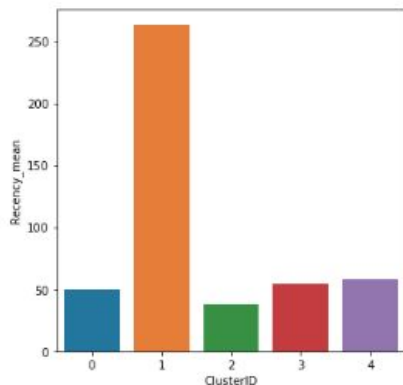
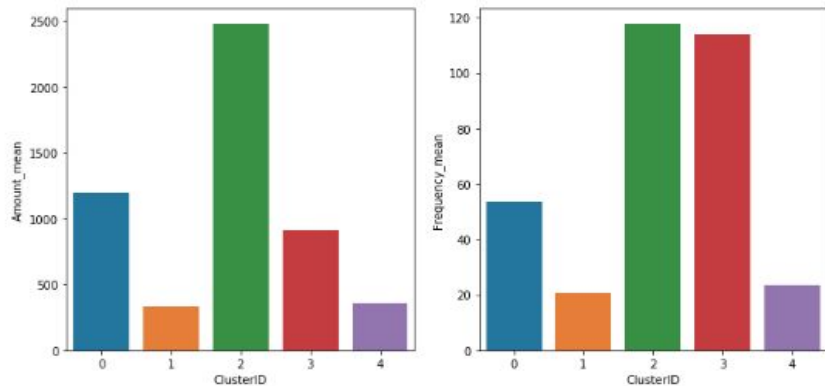
Učenie bez učiteľa - clustering

- Príklad:

▲ InvoiceNo	▲ StockCode	▲ Description	# Quantity	📅 InvoiceDate	# UnitPrice	👤 CustomerID	🌐 Country
25900 unique values	4070 unique values	4224 unique values					
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850	United Kingdom
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850	United Kingdom
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	01-12-2010 08:26	7.65	17850	United Kingdom
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	01-12-2010 08:26	4.25	17850	United Kingdom
536366	22633	HAND WARMER UNION JACK	6	01-12-2010 08:28	1.85	17850	United Kingdom



Učenie bez učiteľa - clustering



Poznámky:

- Triedy (typy zákazníkov) neboli určené dopredu, analýza bola spravená po natrénovaní modelu
- Zhluky boli vytvorené pomocou všetkých kategórií zo vstupu
- Zhlukovanie vyberá (malo by) vyberať tie triedy tak, aby:
 - Vzorky v rámci jedného zhluku si navzájom najpodobnejšie (intratriedna vzdialenosť)
 - Zhluky medzi sebou (intertriedna vzdialenosť) sú si čo najmenej podobné

1. Ako určím podobnosť medzi vzorkami?
2. Aký je správny počet zhlukov?



Meranie vzdialeností

- Opäť potrebujeme číselné reprezentácie pre každý vstup
- Minkowski vzdialenosť:

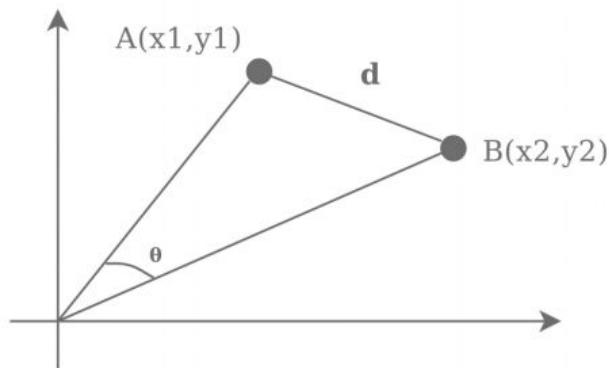
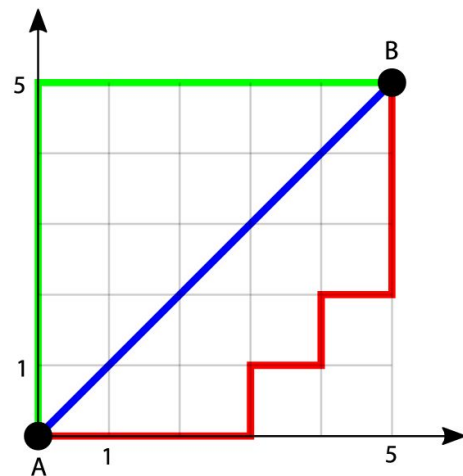
$$d = \left(\sum_{i=1}^n |A_i - B_i|^p \right)^{1/p}$$

- Manhattanská vzdialenosť $p = 1$
- Euklidovská vzdialenosť $p = 2$

- Kosínusová vzdialenosť:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$

- Mahalanobis, Hamming ...



Zhlukovanie

Zdroje obrázkov:
https://eugenfunk.files.wordpress.com/2013/09/clustering_scheme1.png?w=640
<https://www.statisticshowto.com/hierarchical-clustering/>

- Connectivity models

- Hierarchické modely
- Nevhodná pre veľké datasety
- AHC

- Centroid models

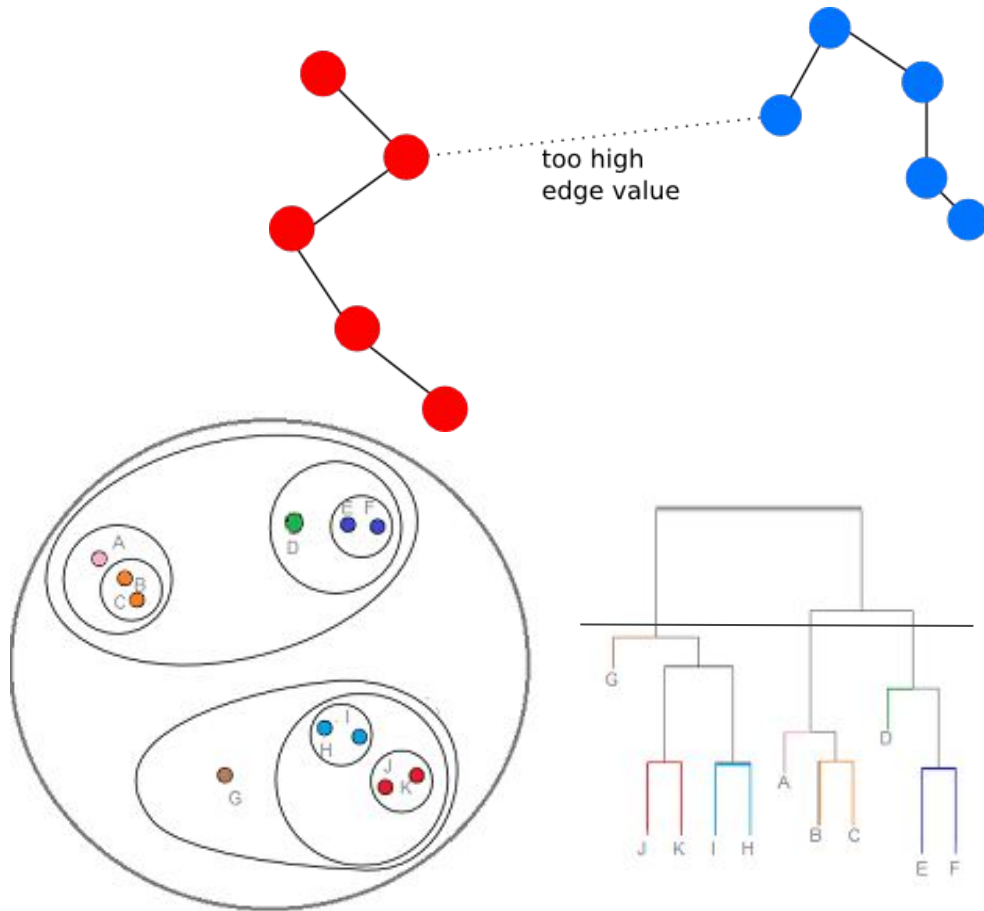
- Počet zhlukov určený predom
- Citlivé na inicializáciu a outliers
- K-means

- Distribution models

- Založené na pravdepodobnosti
- Náchylné na pretrénovanie
- Gaussian Mixture Models

- Density models

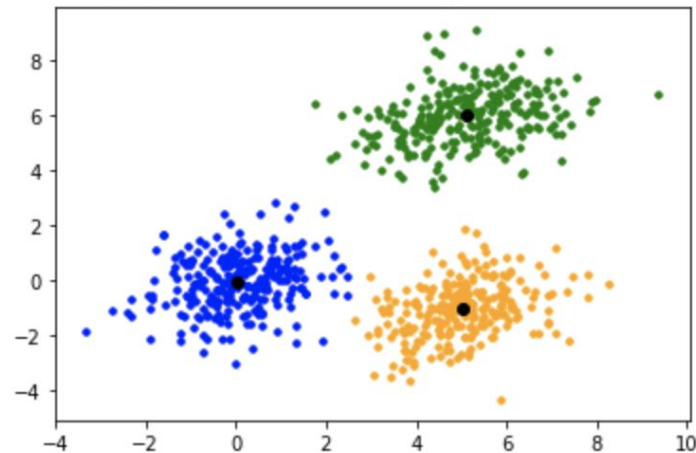
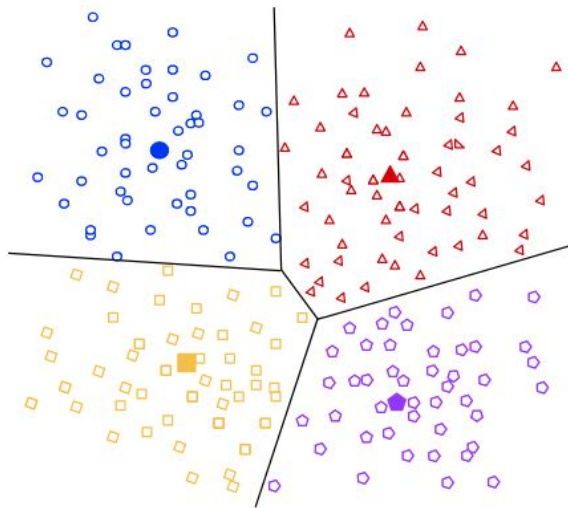
- Hľadajú "husto obsadené" podpriestory
- DBSCAN



Zhlukovanie

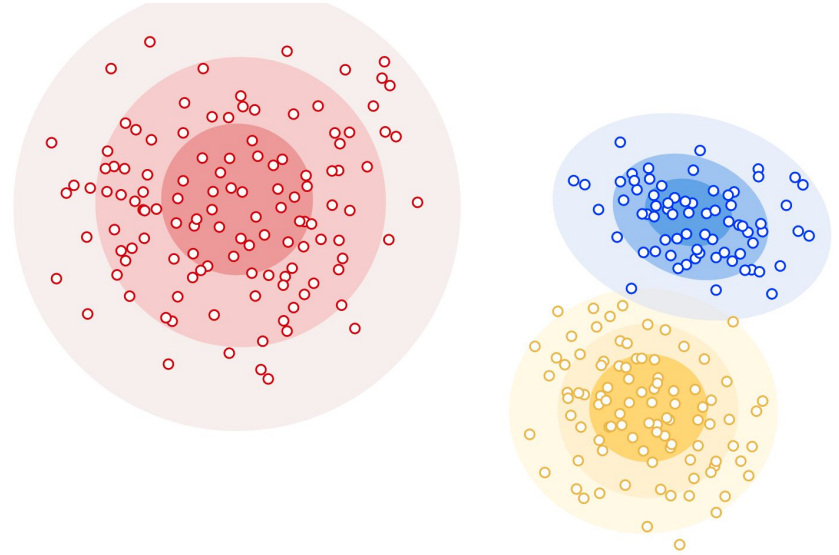
Zdroje obrázkov:
<https://www.geeksforgeeks.org/ml-k-means-algorithm/>
<https://developers.google.com/machine-learning/clustering/clustering-algorithms>

- Connectivity models
 - Hierarchické modely
 - Nevhodná pre veľké datasety
 - Chinese whispers
- **Centroid models**
 - Počet zhukov určený predom
 - Citlivé na inicializáciu a outliers
 - K-means
- Distribution models
 - Založené na pravdepodobnosti
 - Náchylné na pretrénovanie
 - Gaussian Mixture Models
- Density models
 - Hľadajú "husto obsadené" podpriestory
 - DBSCAN



Zhlukovanie

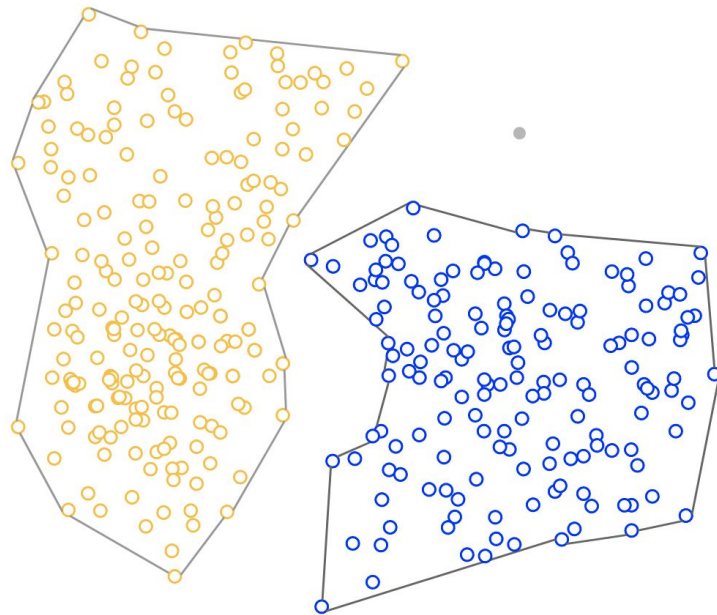
- Connectivity models
 - Hierarchické modely
 - Nevhodná pre veľké datasety
 - Chinese whispers
- Centroid models
 - Počet zhlukov určený predom
 - Citlivé na inicializáciu a outliers
 - K-means
- **Distribution models**
 - Založené na pravdepodobnosti
 - Náchylné na pretrénovanie
 - Gaussian Mixture Models
- Density models
 - Hľadajú “husto obsadené” podprieštory
 - DBSCAN



Zhlukovanie

Zdroje obrázkov:
<https://developers.google.com/machine-learning/clustering/clustering-algorithms>

- Connectivity models
 - Hierarchické modely
 - Nevhodná pre veľké datasety
 - Chinese whispers
- Centroid models
 - Počet zhukov určený predom
 - Citlivé na inicializáciu a outliers
 - K-means
- Distribution models
 - Založené na pravdepodobnosti
 - Náchylné na pretrénovanie
 - Gaussian Mixture Models
- **Density models**
 - Hľadajú “husto obsadené” podpriestory
 - DBSCAN

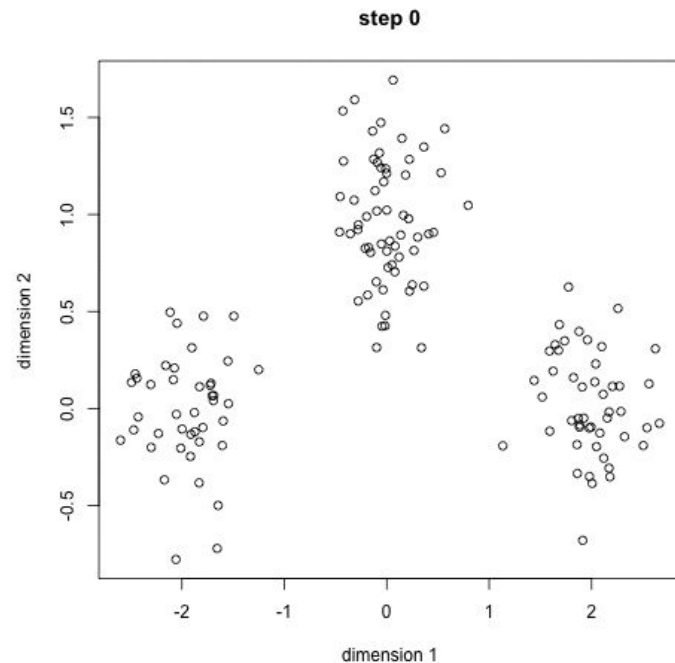


K-means clustering

1. Zvoľme počet clusterov, každému prislúcha jeden centroid zvolený náhodne
2. Každý bod priradíme ku tomu zhľuku, ktorého centroid je mu najbližšie
3. Každému zhľuku vypočítame nový centroid ako priemer bodov priradených zhľuku
4. Opakujeme bod 2-3 do konvergenencie alebo zastavujúcej podmienky

Príklad:

<https://www.naftaliharris.com/blog/visualizing-k-means-clustering/>



DBSCAN

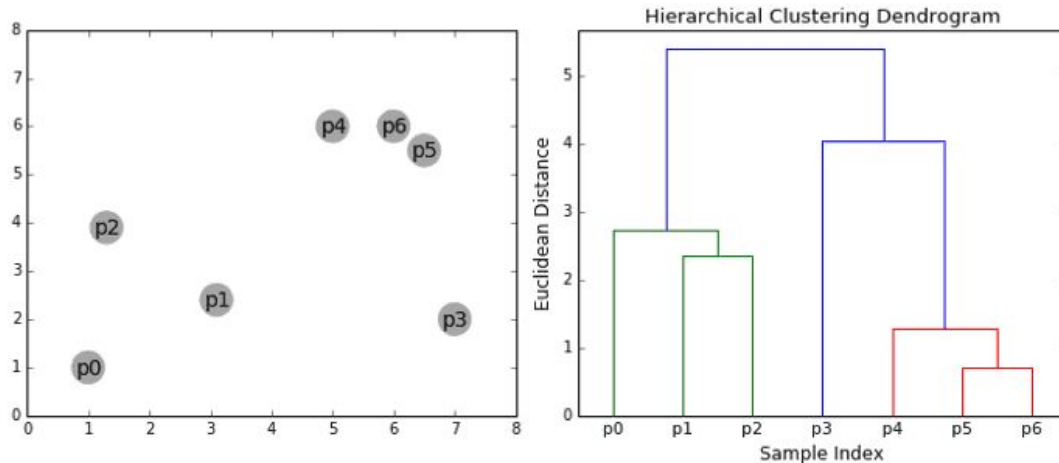
1. Vyberme bod, ktorý ešte nebol navštívený.
2. Ak má v okolí ($d < \text{hyperparameter } \epsilon$) dostatočný počet bodov (hyperparameter minPoints), stáva sa z neho počiatok clustra. Body z okolia pridáme do clustra. Označíme ho ako navštívený.
3. Pre každý nenavštívený bod v clustri, pridáme všetky body z jeho okolia do clustra a označíme ho ako navštívený.
4. Opakujeme bod 3, kým je čo pridať.
5. Opakujeme bod 1-4 do konvergenzie alebo zastavujúcej podmienky.

Príklad: <https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/>



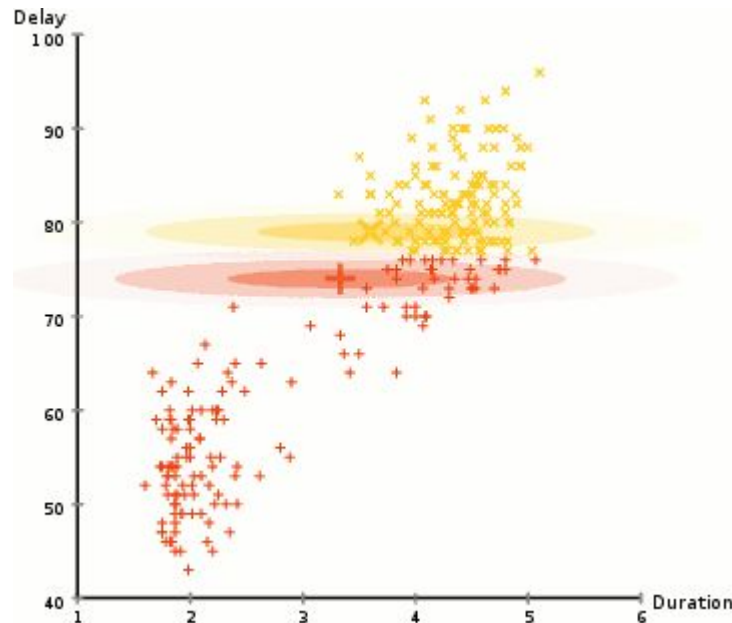
Agglomerative Hierarchical Clustering

1. Každý bod je vlastným zhlukom.
2. Skombinujeme 2 najbližšie zhluky do jedného. (Vyberáme akým spôsobom počítame vzdialenosti medzi zhlukmi).
3. Opakujeme bod 2, kým nie sú všetky body v jednom zhluku.



Gaussian Mixture Models

1. Vyberieme počet zhlukov a náhodne inicializujeme parametre pre Gaussovo rozdelenie pravdepodobnosti
2. Vypočítame pravdepodobnosť, že bod patrí do zhluku
3. Upravíme parametre tak, aby sme maximalizovali pravdepodobnosti, že body patria k zhluku
4. Opakujeme bod 2-3 do konvergenzie alebo zastavujúcej podmienky



Exploratory Data Analysis

- Wikipedia:
 - *In statistics, exploratory data analysis is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.*
- Začiatok práce s datasetom - hľadá vzory, odpovedá na otázky o dátach, odhaľuje skryté vzťahy, pomáha pri *feature engineering*
- Príklady:
 - <https://www.kaggle.com/lmorgan95/r-suicide-rates-in-depth-stats-insights>
 - <https://www.kaggle.com/donyoe/exploring-youtube-trending-statistics-eda>
 - <https://www.kaggle.com/adhok93/eda-with-plotly>
 - <https://www.kaggle.com/xvivancos/eda-tweets-during-cavaliers-vs-warriors>
 - <https://www.kaggle.com/danilodiogo/google-play-store-eda-plotting-with-highcharts>
 - <https://www.kaggle.com/xvivancos/eda-the-cure-discography>
 - <https://www.kaggle.com/erikbruin/gun-violence-in-the-us-eda-and-rshiny-app>
 - <https://www.kaggle.com/lucian18/matching-loans-with-poverty-problems>

