



TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií ■

Vizualizace dat

František Kynych
7. 10. 2021 | MVD





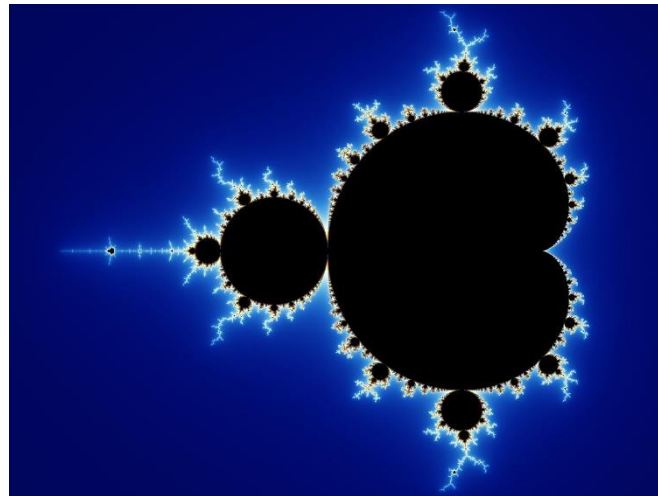
TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky
a mezioborových studií ■

Část I.: Úvod, dělení dat



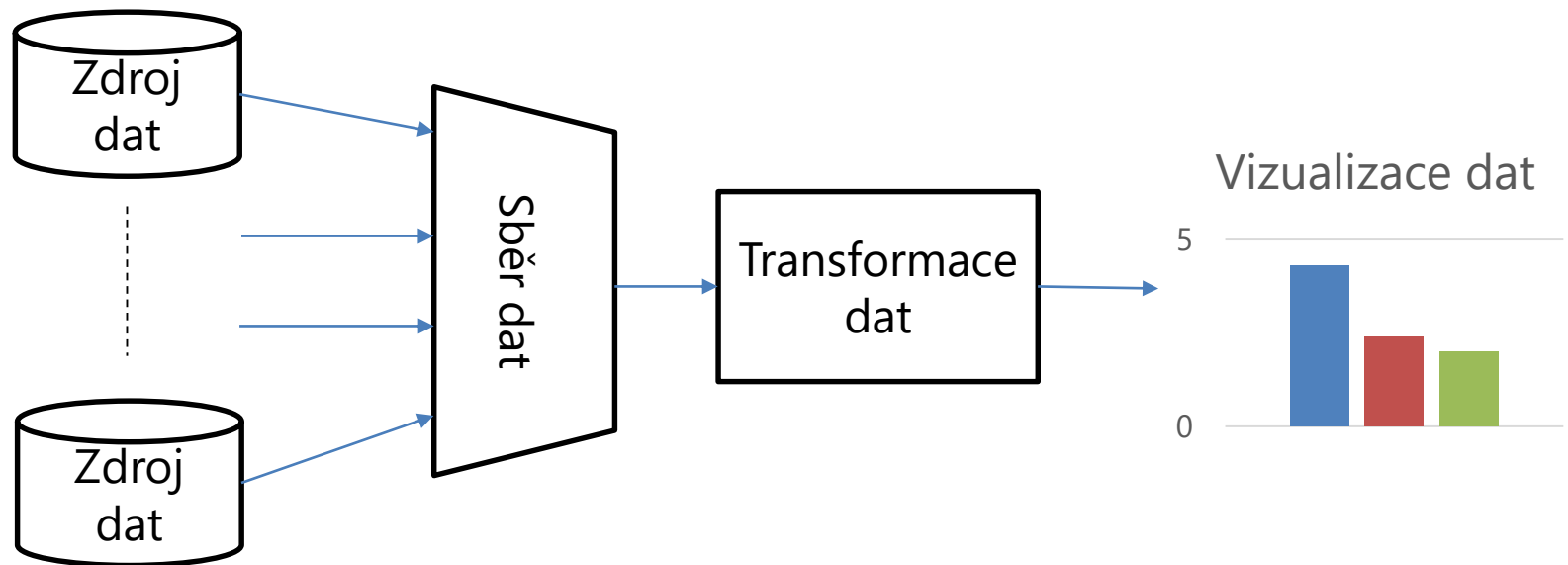
Důvod vizualizace dat

- Efektivní přenos informace uživateli
- Získání přehledu nad velkým množstvím informací
- Z vizualizace můžeme rychleji zjistit, co se nám snaží data říct
- Důležitý prvek zdravotnictví, business intelligence (BI), vzdělávání a mnoha dalších odvětví



Zdroj: <http://edu.techmania.cz/encyklopedie/matematika/geometrie/fraktaly>

System vizualizace dat



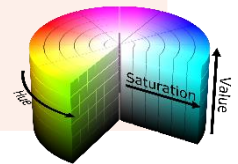
Druhy dat

Základní dělení typů dat:

- **Kvalitativní**
 - Kategoriální
 - Lze je zařadit do kategorií, ale nelze je kvantifikovat
- **Kvantitativní**
 - Numerická
 - Lze je charakterizovat číselnou hodnotou

Druhy dat

Hodnoty	Diskrétní	Spojité
Seřazené	Ordinální <ul style="list-style-type: none"> - S, M, L, XL Kvantitativní <ul style="list-style-type: none"> - 1, 2, 3 	Intervalové <ul style="list-style-type: none"> - teplota - nadmořská výška
Neseřazené (nelze je porovnat)	Nominální <ul style="list-style-type: none"> - geometrické tvary Kategorie <ul style="list-style-type: none"> - národnost 	Opakující se (cyklické) <ul style="list-style-type: none"> - Hue v HSV modelu



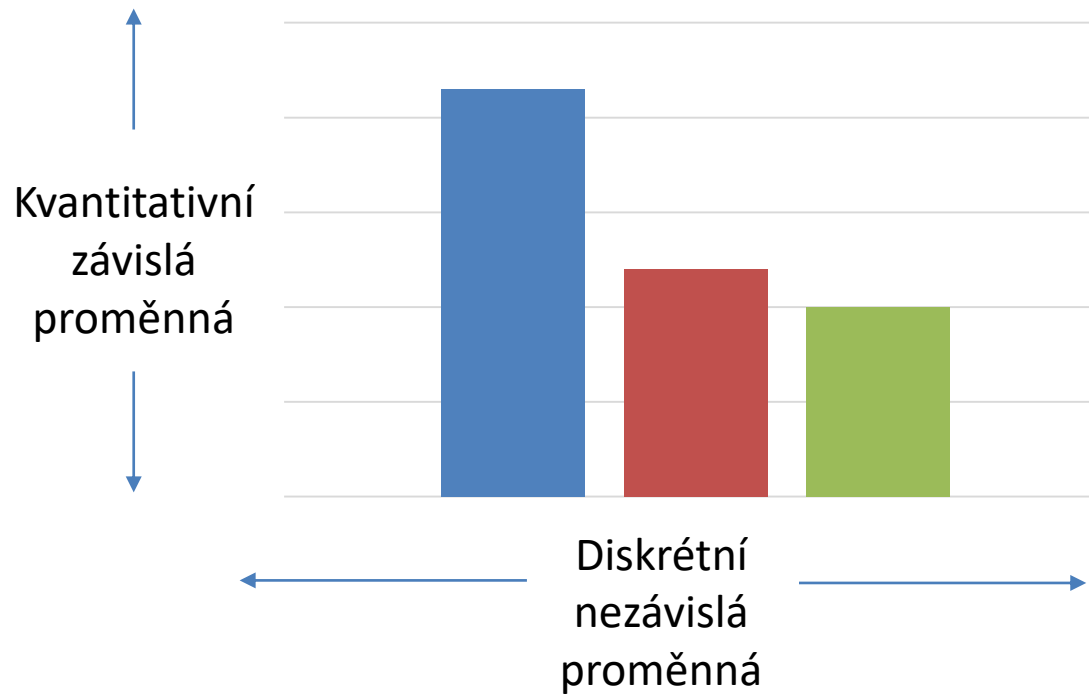
Závislé a nezávislé proměnné

- Nezávislá proměnná
 - Ta, se kterou manipulujeme
- Závislá proměnná
 - Měřená proměnná
- Účel
 - Pomoc při rozhodování o tom, jak nejlépe zobrazit data
- Příklad
 - Vztah klíč -> hodnota
 - Nezávislá prom. -> závislá prom.



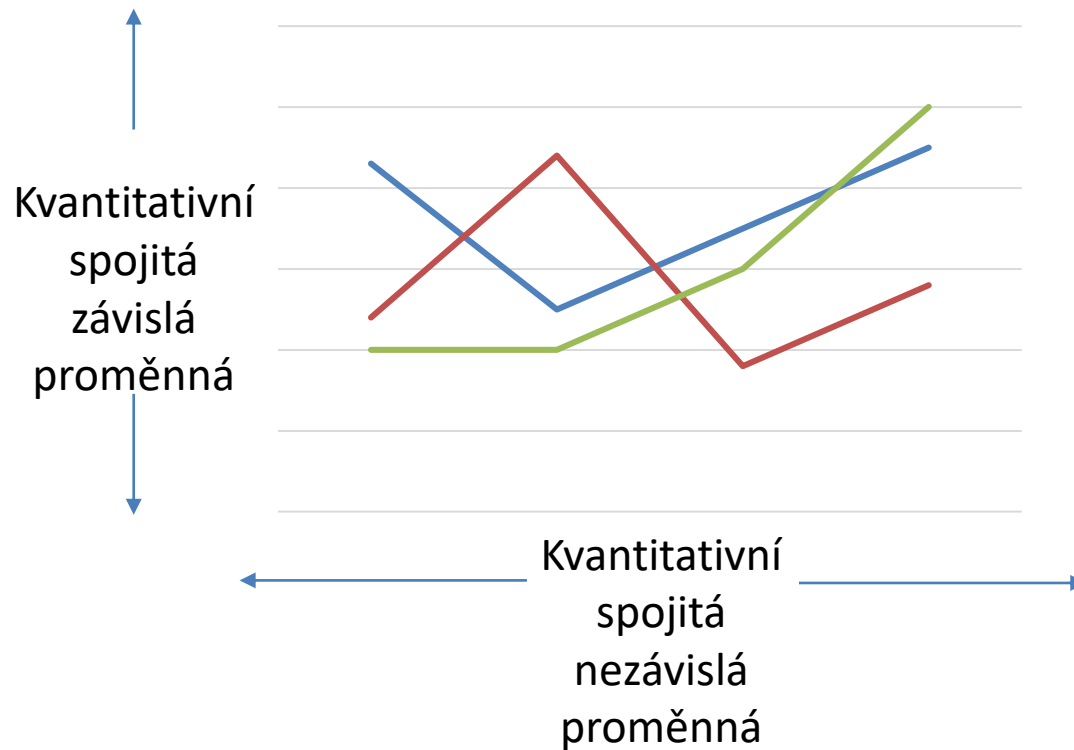
Část II.: Základní možnosti vizualizace

Sloupcový graf

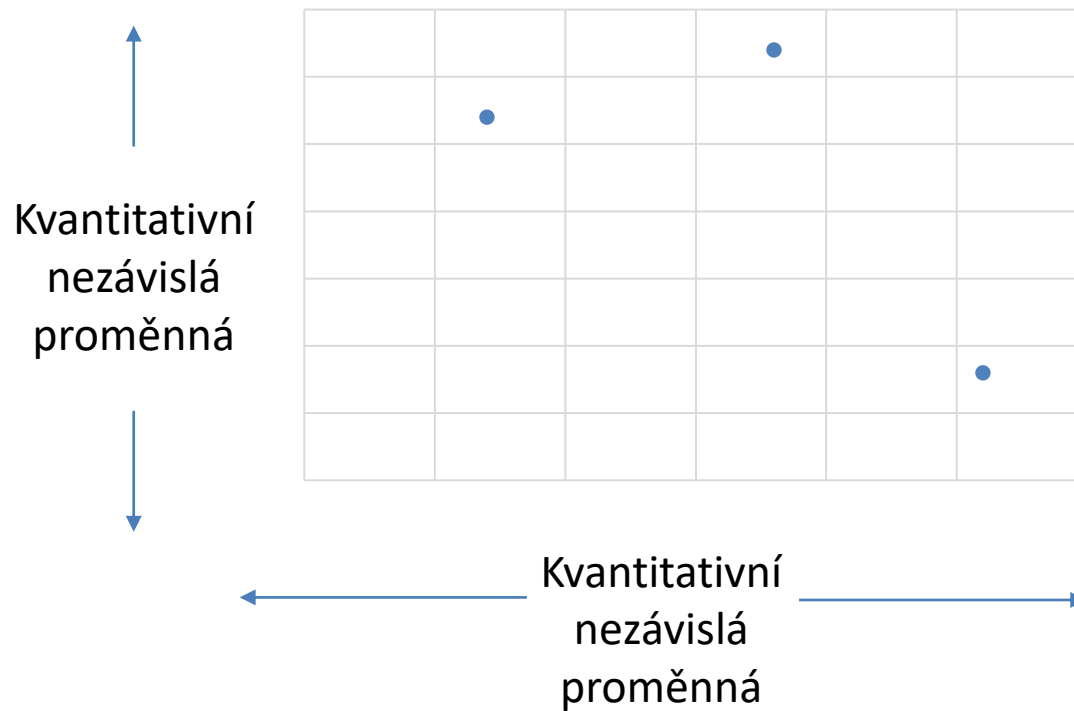


Liniový graf

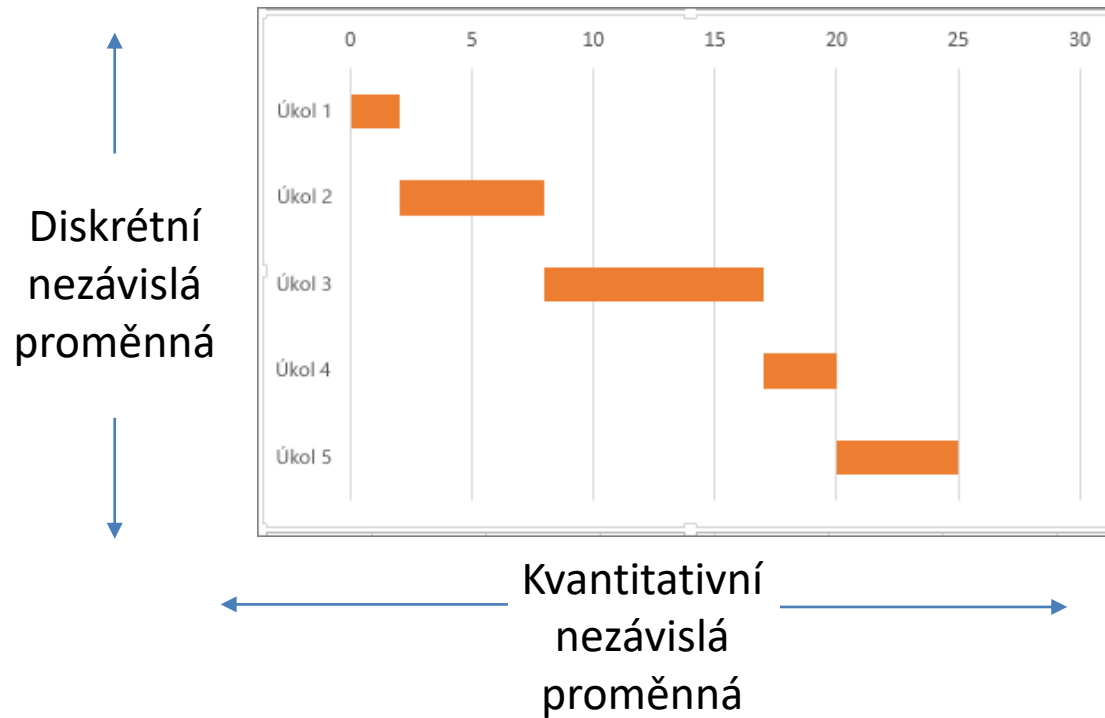
- Také čárový nebo spojnicový graf



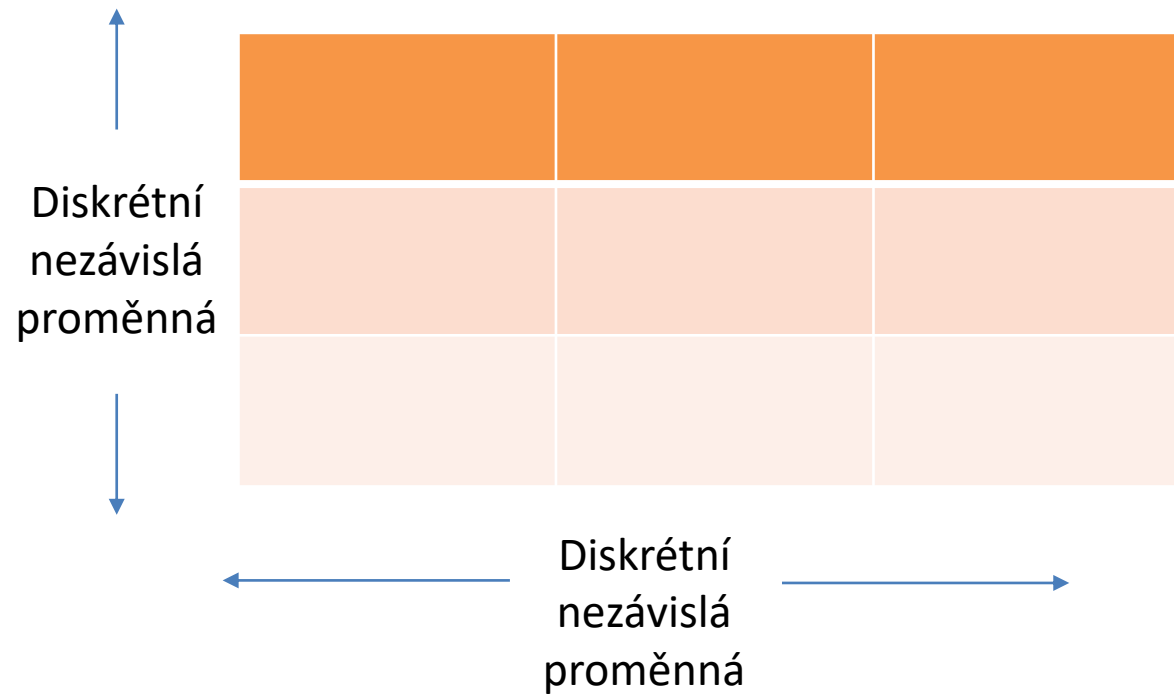
Bodový graf



Ganttův diagram



Tabulka



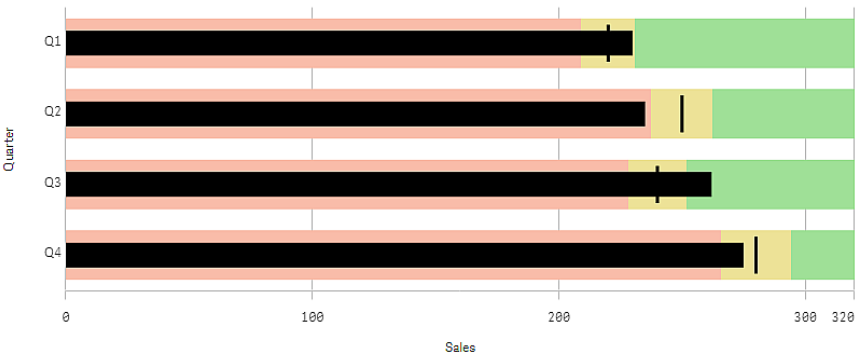
Co použít?

		X	
		Nezávislá	
		Diskrétní	Spojitá
Y	Závislá	Kvantitativní Spojitá	Sloupcový graf Liniový graf
		Kvantitativní Diskrétní	Sloupcový graf Sloupcový graf
	Nezávislá	Kvantitativní Spojitá	Ganttův diagram Bodový graf
		Diskrétní	Tabulka Ganttův diagram

Část III.: Další možnosti vizualizace a kdy je využít

Porovnání hodnot

Bullet chart



Zdroj: help.qlik.com/en-US/sense/June2020/Subsystems/Hub/Content/Resources/Images/ex_gen_bullet_chart.png

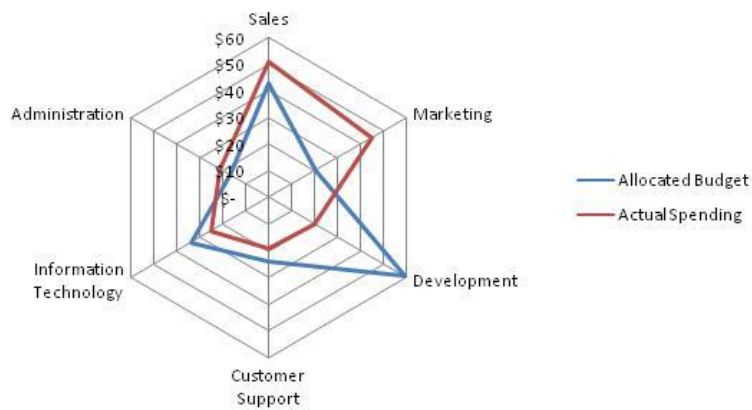
Waterfall chart



Zdroj: https://en.wikipedia.org/wiki/Waterfall_chart

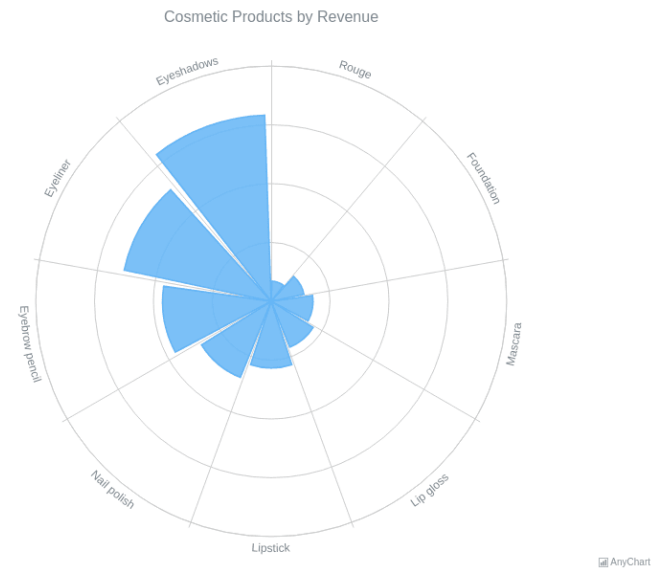
Porovnání hodnot

Radar chart



Zdroj: https://en.wikipedia.org/wiki/Radar_chart

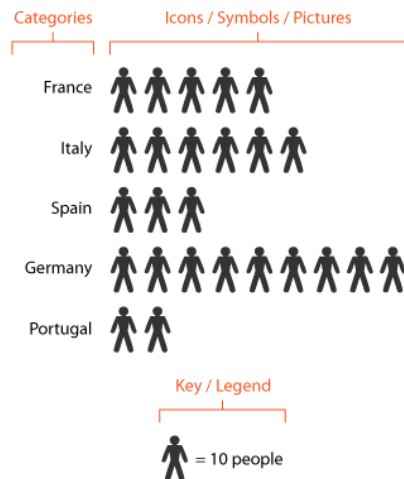
Polar chart



Zdroj: https://www.anychart.com/products/anychart/gallery/Polar_Charts/

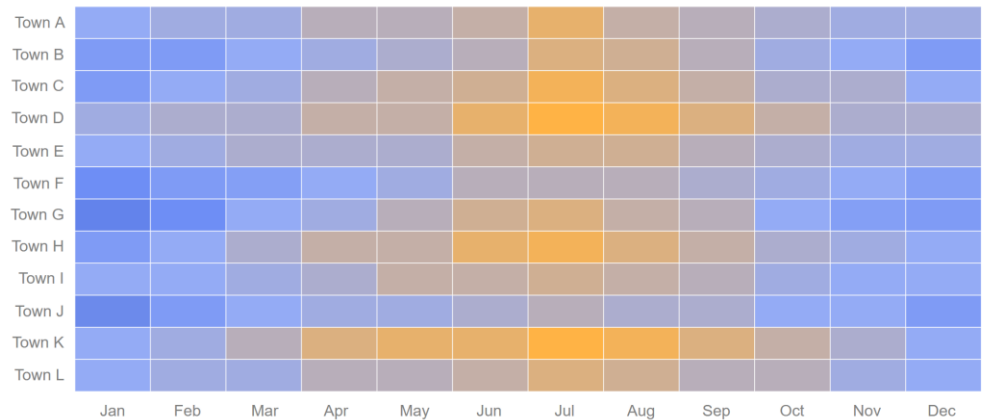
Porovnání hodnot

Pictogram chart



Zdroj: <https://dataforvisualization.com/charts/pictogram-chart/>

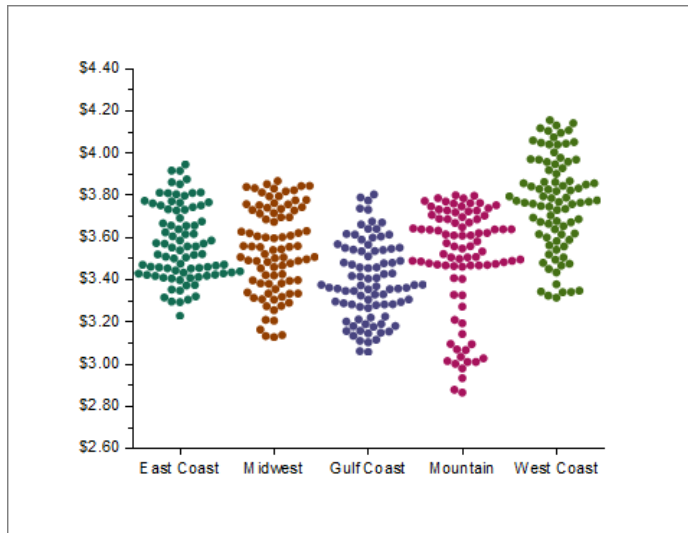
Heat map



Zdroj: <https://datavizcatalogue.com/methods/heatmap.html>

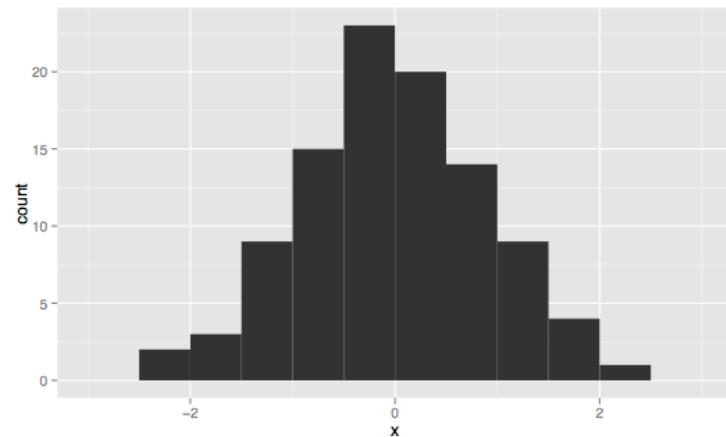
Porovnání rozložení

Beeswarm plot



Zdroj: <https://www.originlab.com/doc/Origin-Help/Beeswarm-Plot>

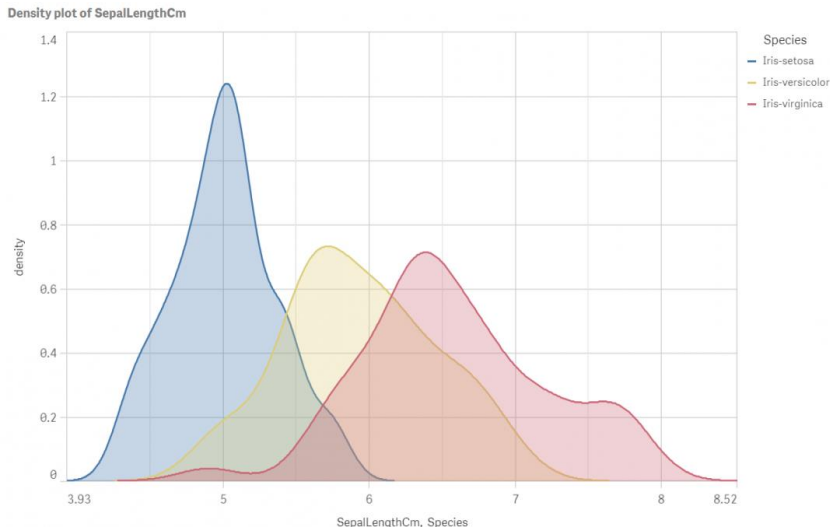
Histogram



Zdroj: <https://en.wikipedia.org/wiki/Histogram>

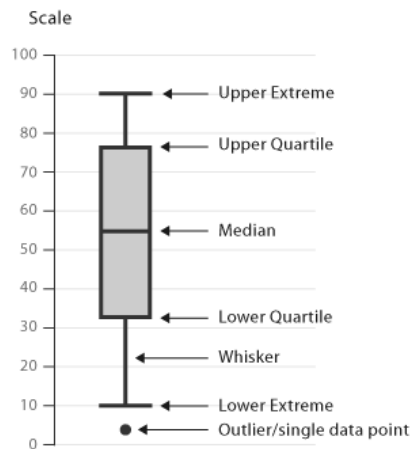
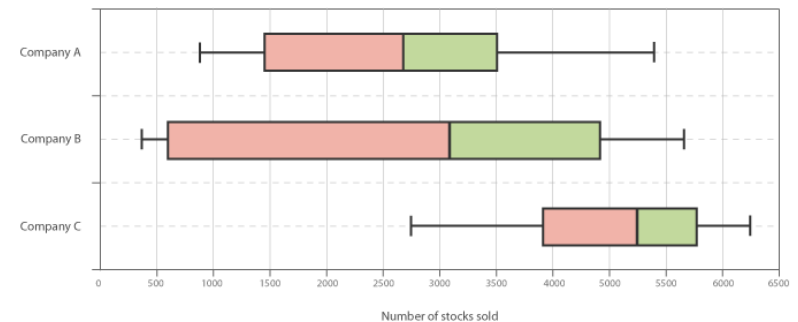
Vizualizace rozložení

Density plot



Zdroj: https://datavizcatalogue.com/methods/density_plot.html

Box and Whisker Plot



Zdroj: https://datavizcatalogue.com/methods/box_plot.html

Vizualizace hierarchií

Treemap

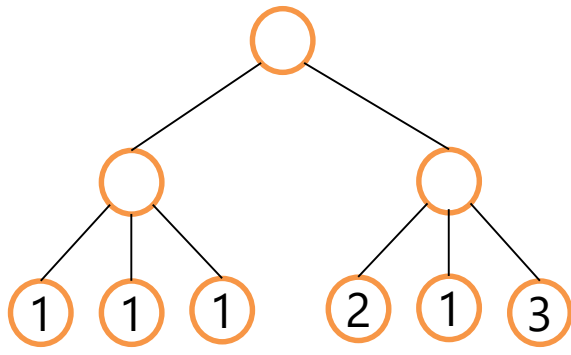


Zdroj: <https://finviz.com/map.ashx>



Vizualizace hierarchií

Jak vzniká treemap

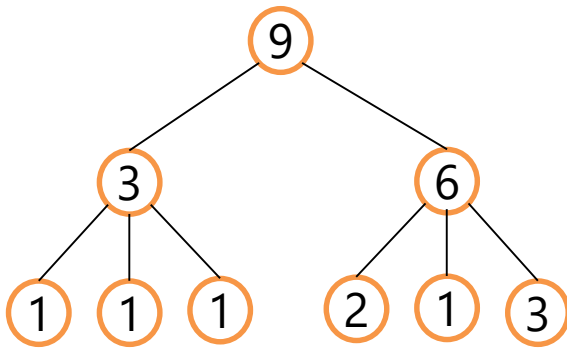


Hodnoty pro 2 různé kategorie



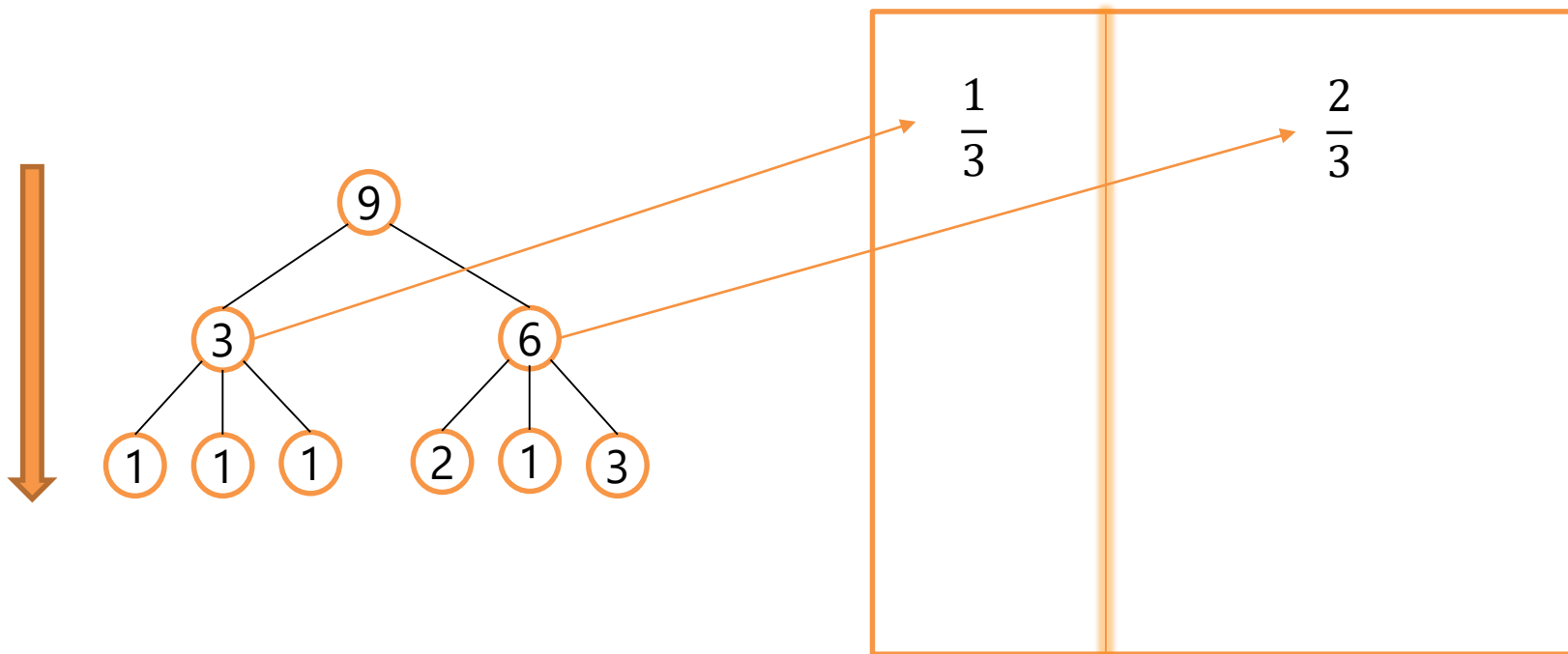
Vizualizace hierarchií

Jak vzniká treemap



Vizualizace hierarchií

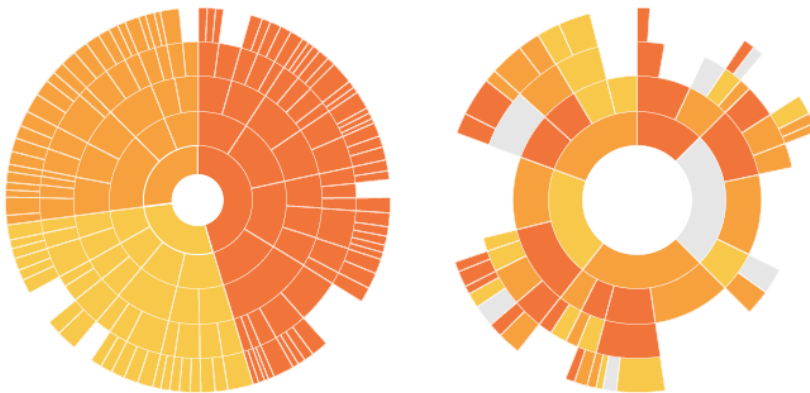
Jak vzniká treemap



- Stejným způsobem pokračujeme dále (levou část rozdělíme na třetiny, ...)

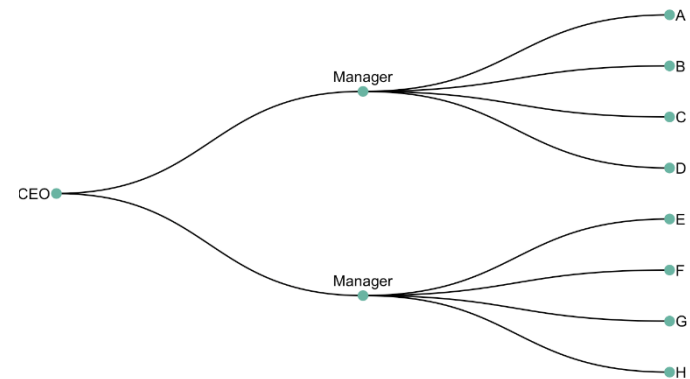
Vizualizace hierarchií

Sunburst chart



Zdroj: https://datavizcatalogue.com/methods/sunburst_diagram.html

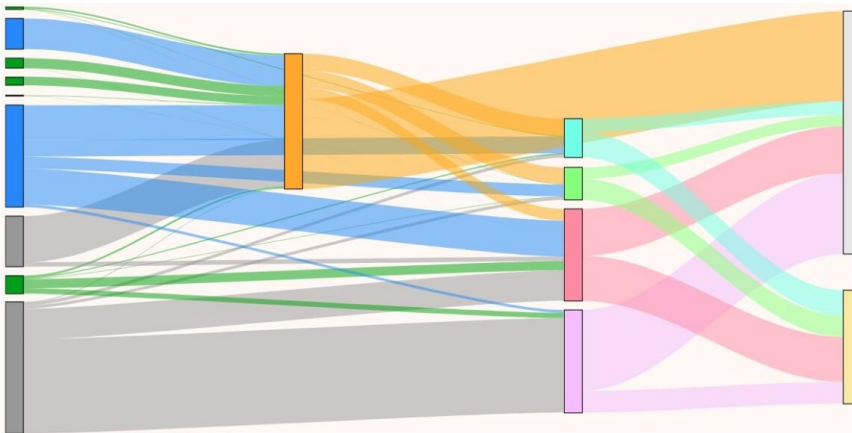
Dendrogram



Zdroj: <https://www.data-to-viz.com/graph/dendrogram.html>

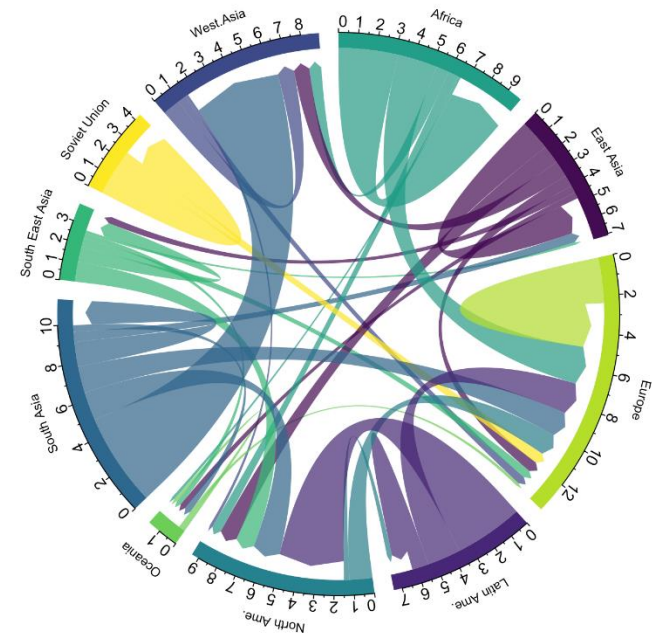
Vizualizace propojení

Sankey diagram



Zdroj: <https://www.highcharts.com/blog/tutorials/what-is-a-sankey-diagram/>

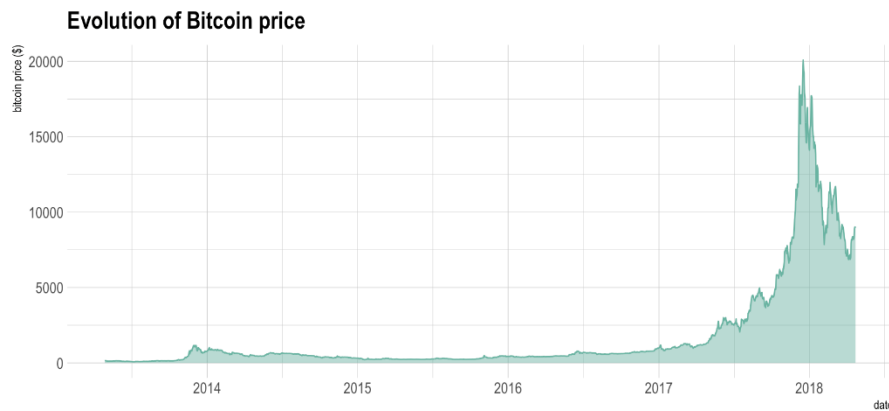
Chord diagram



Zdroj: <https://www.data-to-viz.com/graph/chord.html>

Vizualizace trendu

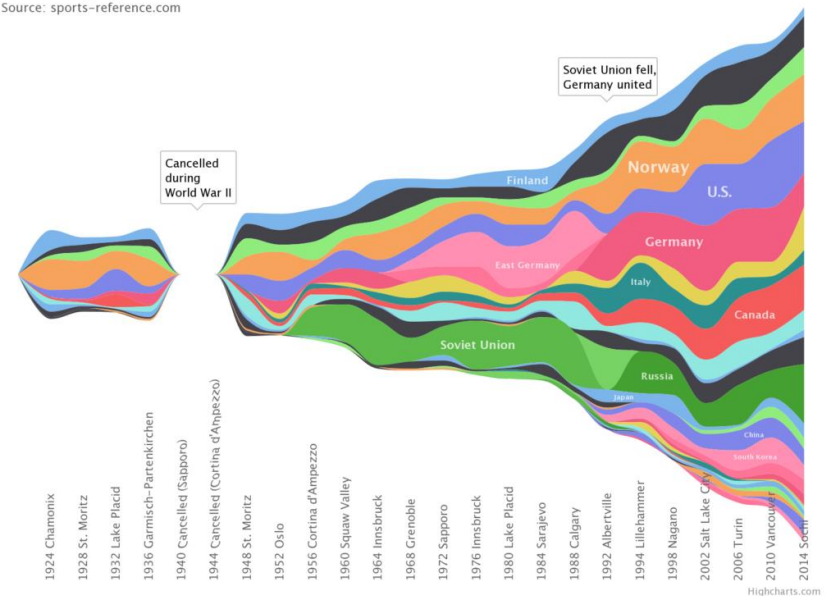
Area chart



Zdroj: <https://www.data-to-viz.com/graph/area.html>

Stream chart

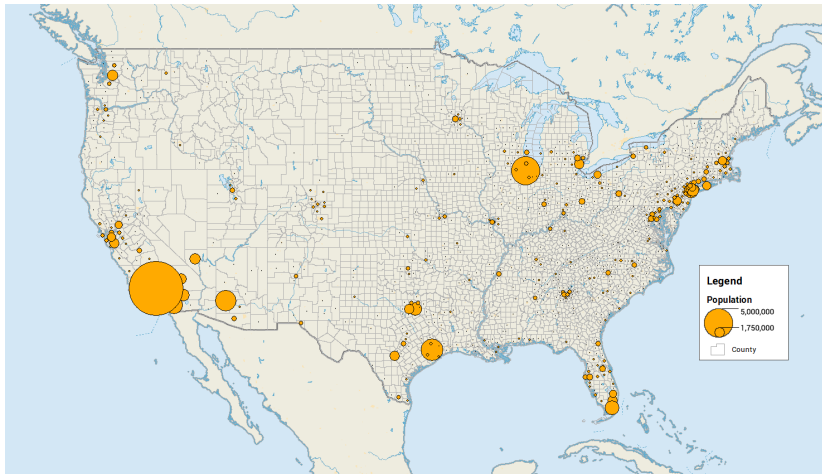
Winter Olympic Medal Wins
Source: sports-reference.com



Zdroj: <https://dataforvisualization.com/charts/stream-chart/>

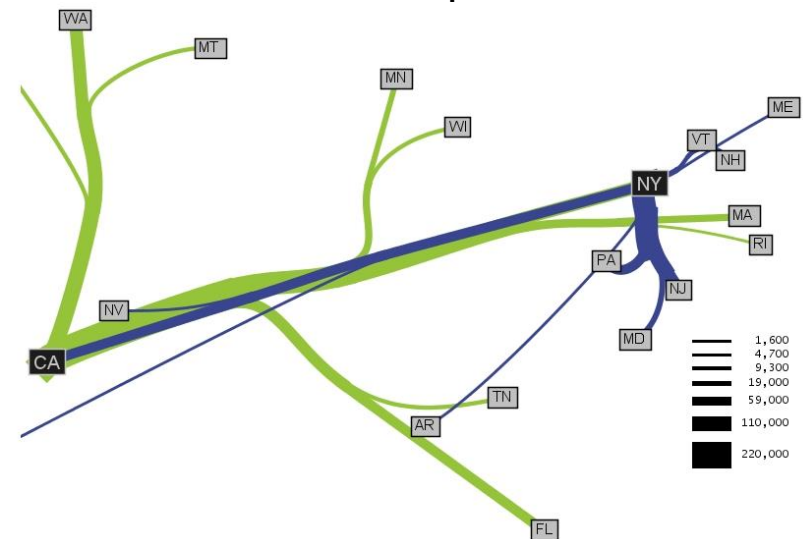
Geografické vizualizace

Area chart



Zdroj: <https://gisgeography.com/dot-distribution-graduated-symbols-proportional-symbol-maps/>

Flow map



Zdroj: https://graphics.stanford.edu/papers/flow_map_layout/

Část IV:

Redukce dimenzionality dat pro vizualizaci

PCA – opakování z USU

- Principal Component Analysis

PCA – opakování z USU

- Principal Component Analysis
 - Slouží k dekorelaci příznaků nebo k redukci počtu příznaků
1. Normalizujeme data \mathbf{X}
 2. Vypočteme kovarianční matici dat $\Sigma_{\mathbf{X}}$
 3. Vypočteme vlastní čísla λ a vlastní vektory \mathbf{v} kovarianční matice $\Sigma_{\mathbf{X}}$
 4. Vybereme \mathbf{n} vlastních vektorů příslušejících \mathbf{n} největším vlastním číslům (vznikne menší matice \mathbf{V} obsahující vlastní vektory)
 5. Transformujeme data do nižší dimenze
$$\mathbf{Z} = \mathbf{XV}$$

t-SNE

- t-distributed Stochastic Neighbor Embedding
- Často využíváno pro vizualizaci dat
 - Redukce dat do 2 nebo 3 dimenzí
- Dokáže najít i nelineární vztahy mezi daty
 - Často najde strukturu tam, kde jiné algoritmy ne
- Intuice
 - Pokud hodnoty ve vyšší dimenzi spadají do jednoho clusteru
-> měly by i v nižší dimenzi
- Iterativní algoritmus

t-SNE algoritmus

1. Výpočet pravděpodobnostního rozdělení hodnot
 - S jakou pravděpodobností jsou dva body sousedé

$$p_{j|i} = \frac{\exp(\frac{\|x_i - x_j\|^2}{2\sigma_i^2})}{\sum_{k \neq i} \exp(\frac{\|x_i - x_k\|^2}{2\sigma_i^2})}$$

2. Náhodná projekce dat do nižší dimenze a spočteme pro ně studentovo p. rozdělení
3. Minimalizace KL divergence (jak moc se dvě p. rozdělení liší)

t-SNE parametry

- Dimenze výsledného prostoru
- Perplexita
 - Použito pro výběr variance Gaussova rozdělení
 - Často vede k velkým změnám ve výsledné vizualizaci
 - Intuice
 - Počet sousedů, kteří mají vliv na daný bod
 - Doporučená hodnota 5 až 50
- Learning rate, počet iterací, ...



Část V: Knihovny a nástroje pro vizualizaci dat

Nejpopulárnější Python knihovny

1. Matplotlib

- Knihovna inspirovaná Matlab prostředím
- Jednoduchá manipulace a úprava většiny částí grafu
- Velké množství knihoven, které rozšiřují Matplotlib

2. Plotly

- S menším množstvím řádků se vytvoří esteticky příjemnější grafy
- Postaveno na původním Plotly.js
 - Možnost interaktivních webových vizualizací

3. Seaborn

- Rozšiřuje Matplotlib knihovnu
 - Graficky lepší grafy
 - Lepší spolupráce s knihovnou Pandas

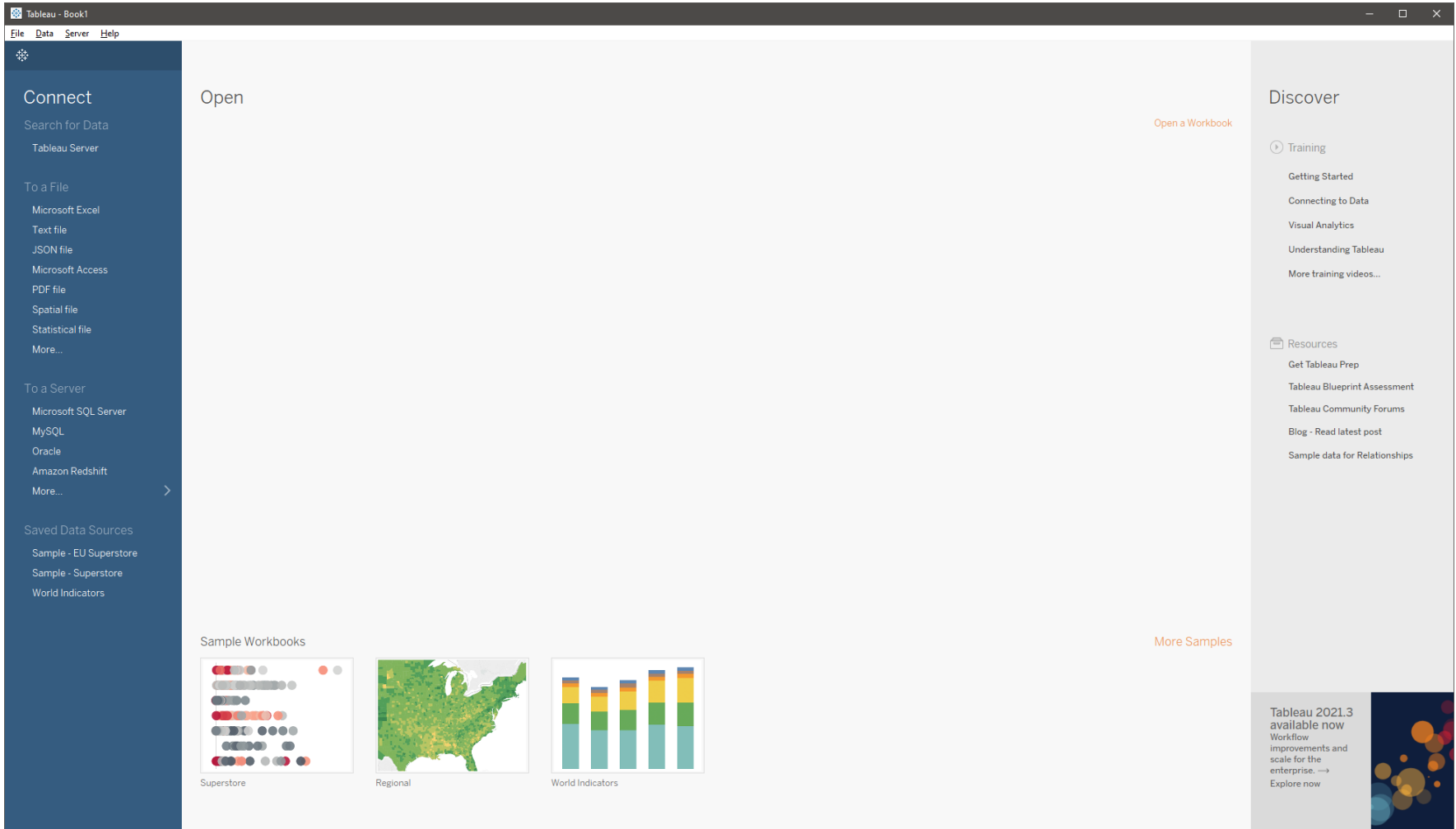
Nástroje pro vizualizaci dat

- Nástroje umožňující jednoduchou vizualizaci a analýzu dat
- Možnost propojení s tabulkami, databázemi nebo cloudy
- Jednodušší pro práci s daty
 - Často bez potřeby programování
 - Drag & drop
- Vytvoření interaktivních dashboardů
- Power BI, Tableau, Qlik Sense, IBM Cognos, ...

Tableau

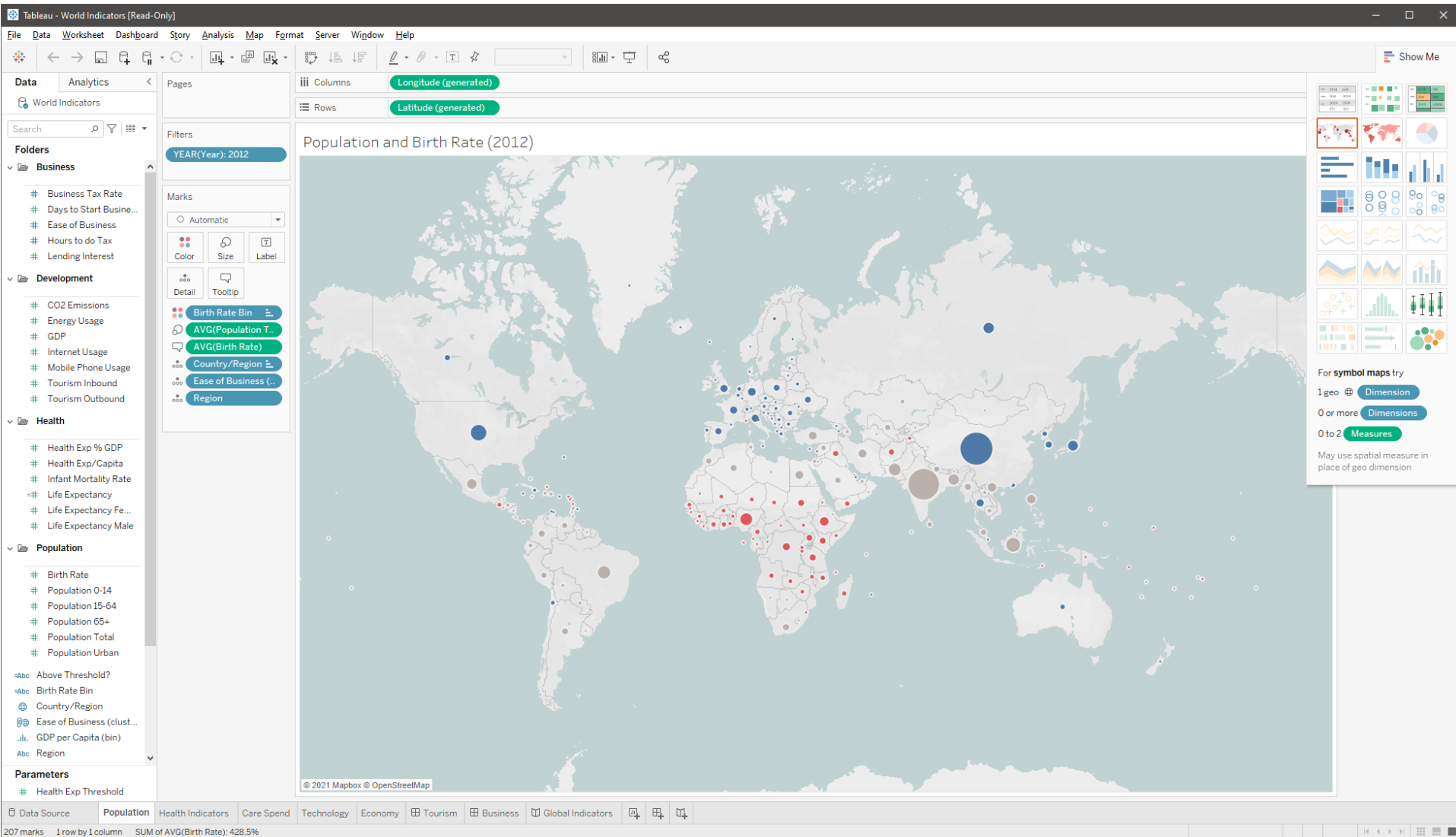
- Jeden ze zástupců aplikací pro vizualizaci dat v business intelligence
- Zvládne pracovat s velkým množstvím dat
 - Lokálně / na cloudu
 - Ostatním platformy jsou často limitovány (např. Power BI)
- Pomáhá čistit a kombinovat data pro analýzu
- Možnost vytvoření reportů, dashboardů nebo stories
- Studentská licence

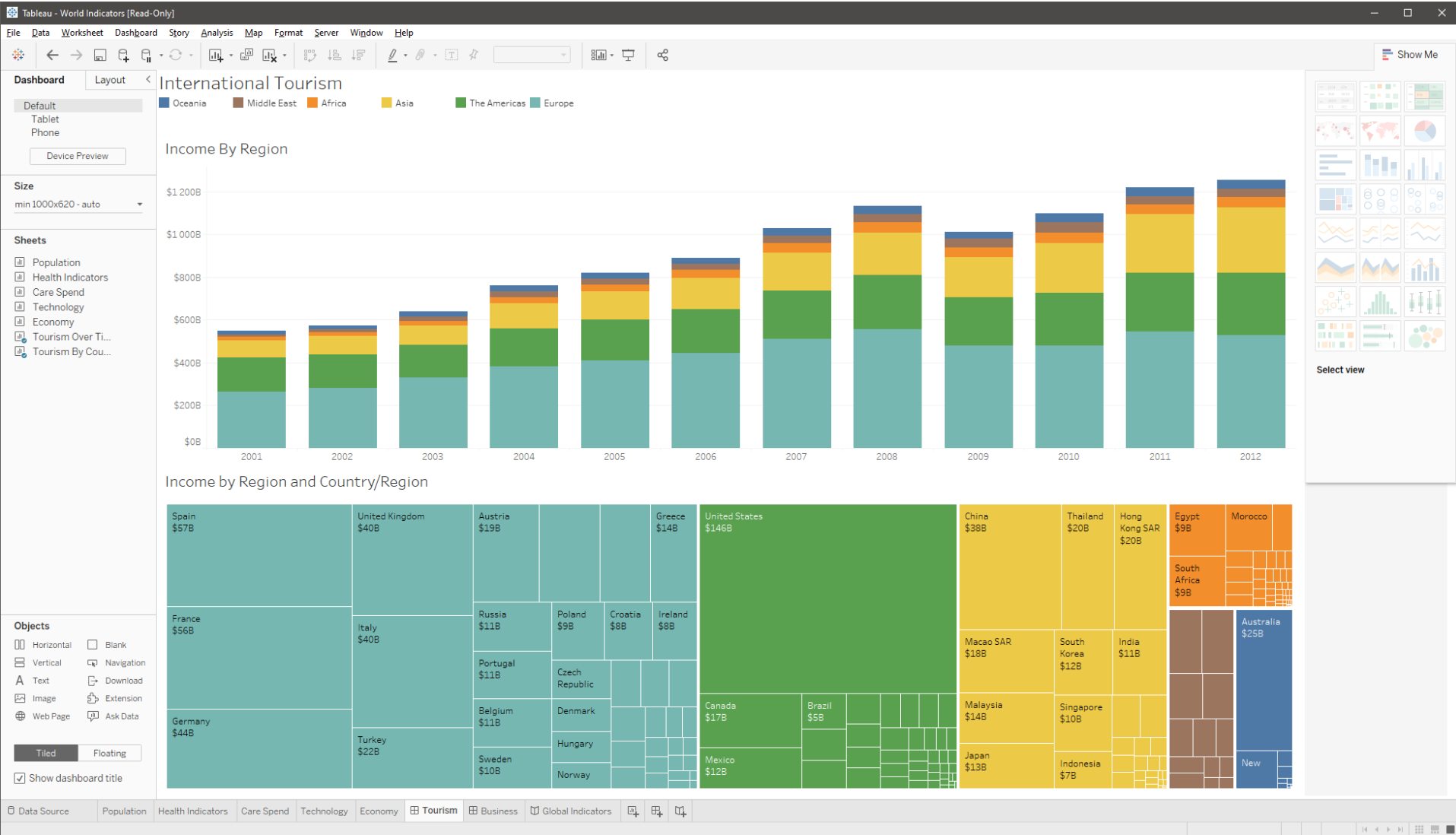
Tableau



The screenshot displays the Tableau Desktop interface. The top menu bar includes 'File', 'Data', 'Server', and 'Help'. The left sidebar is titled 'Connect' and contains sections for 'Search for Data' (Tableau Server), 'To a File' (Microsoft Excel, Text file, JSON file, Microsoft Access, PDF file, Spatial file, Statistical file, More...), 'To a Server' (Microsoft SQL Server, MySQL, Oracle, Amazon Redshift, More...), and 'Saved Data Sources' (Sample - EU Superstore, Sample - Superstore, World Indicators). The main workspace is titled 'Open' and features a large 'Open a Workbook' button. Below this, there are three sample workbook thumbnails: 'Superstore' (a bar chart), 'Regional' (a map of the United States), and 'World Indicators' (a stacked bar chart). A 'More Samples' link is visible. The right sidebar is titled 'Discover' and includes sections for 'Training' (Getting Started, Connecting to Data, Visual Analytics, Understanding Tableau, More training videos...) and 'Resources' (Get Tableau Prep, Tableau Blueprint Assessment, Tableau Community Forums, Blog - Read latest post, Sample data for Relationships). At the bottom right, there is a promotional banner for 'Tableau 2021.3 available now' with a link to 'Explore now'.







Schneiderman's mantra

- Organizační principy pro vytváření vizualizačních systémů
 - Lze uplatnit i na jiné úlohy
- 1. Overview First
 - Poskytnou celistvý pohled na data
- 2. Zoom and Filter
 - Přiblížení k bodu zájmu poskytne detailnější pohled
- 3. Details on Demand
 - Např. zobrazení tooltipu



Užitečná literatura / kurzy

- KIRK, Andy. *Data visualisation: a handbook for data driven design*. 2nd Edition. Los Angeles: SAGE, [2019]. ISBN 978-1-5264-6892-5.
- [t-SNE vliv parametrů](#)

