Univerzita Palackého v Olomouci
Přírodovědecká fakulta
Katedra geoinformatiky

orange
DATA MINING

CVIČEBNICE

Příklady k samostatnému procvičování v Orange

Zdena DOBEŠOVÁ

Olomouc 2025

With the support of the Erasmus+ Programme of the European Union

Tato publikace vznikla s podporou Erasmus+ Program, Jean Monnet Module.

Project No. 620791-EPP-1-2020-1-CZ-EPPJMO-MODULE UrbanDM - Data mining and analyzing of urban structures as contribution to European Union studies.

Podpora Evropské komise při tvorbě této publikace nepředstavuje souhlas s obsahem, který odráží pouze názory autorů, a Komise nemůže být zodpovědná za jakékoliv využití informací obsažených v této publikaci.

Za návrh šablony dokumentu děkuji Mgr. Jakubovi Koníčkovi.

Autorem loga a ilustrací Orange je Agnieszka Rovšnik.

OBSAH

1 ÚVODNÍ INFORMACE

Tato cvičebnice navazuje na knihu ORANGE, Praktický návod do cvičení předmětu Data Mining (Dobešová 2022). V první části cvičebnice seznamuje se základy práce s databází Eurostat a Urban Atlas. V další části jsou uvedeny příklady, které slouží k samostatnému opakování a procvičování některých metod uvedených v praktickém návodu pro Orange. Data ve cvičeních jsou čerpána ze statistické databáze Eurostat, která je statistickou databází Evropské Unie (EU).

Ve cvičení si studenti sestaví vlastní workflow podle pokynů nebo obrázků. Každá úloha obsahuje seznam úkolů k vypracování. Kromě aplikace metod by se studenti měli zejména věnovat pečlivé intepretaci výsledků a snažit se odhalovat jevy a souvislosti v datech. Zjištěné poznatky je dobré zapisovat formou odpovědí k jednotlivým úkolům.

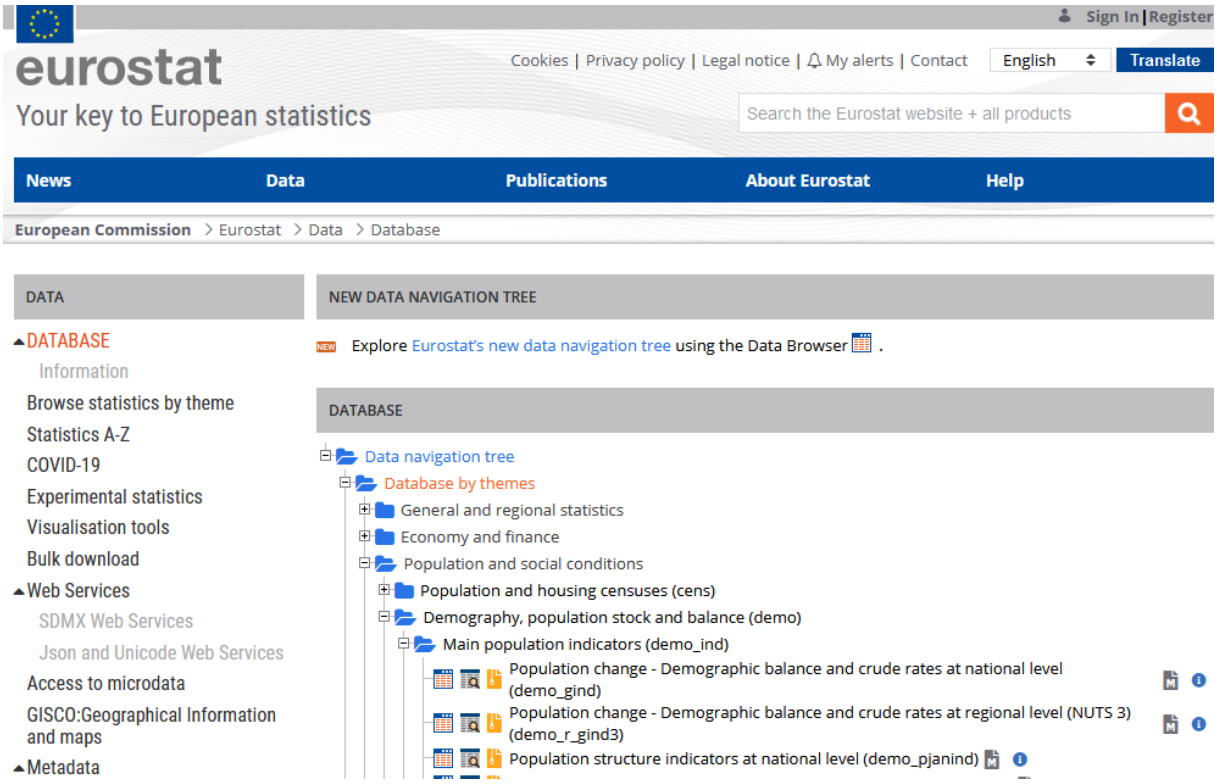
Data používaná v této učebnici jsou dostupná ke stažení na webu **Dobesova.upol.cz/Orange** v souboru **DataUrbanDM.zip**. Doporučujeme studentům stažení aktuálních dat z databáze Eurostat, pokud jsou dostupná.

2 PRÁCE S DATABÁZÍ EUROSTAT

Databáze Eurostat se veřejně dostupná na adrese https://ec.europa.eu/eurostat/data/database.

V databázi je možné vyhledávat v navigačním stromě podle témat, podle politik EU, průřezových témat, nebo jsou nabízeny právě aktualizovaná data (Obr. 1). K dispozici je i nové navigační okno (Obr. 3).

Obr. 1 Rozhraní databáze Eurostat



Obr. 2 Navigační strom a část detailních témat o populaci

Obr. 3 Nové navigační rozhraní rozdělené podle témat

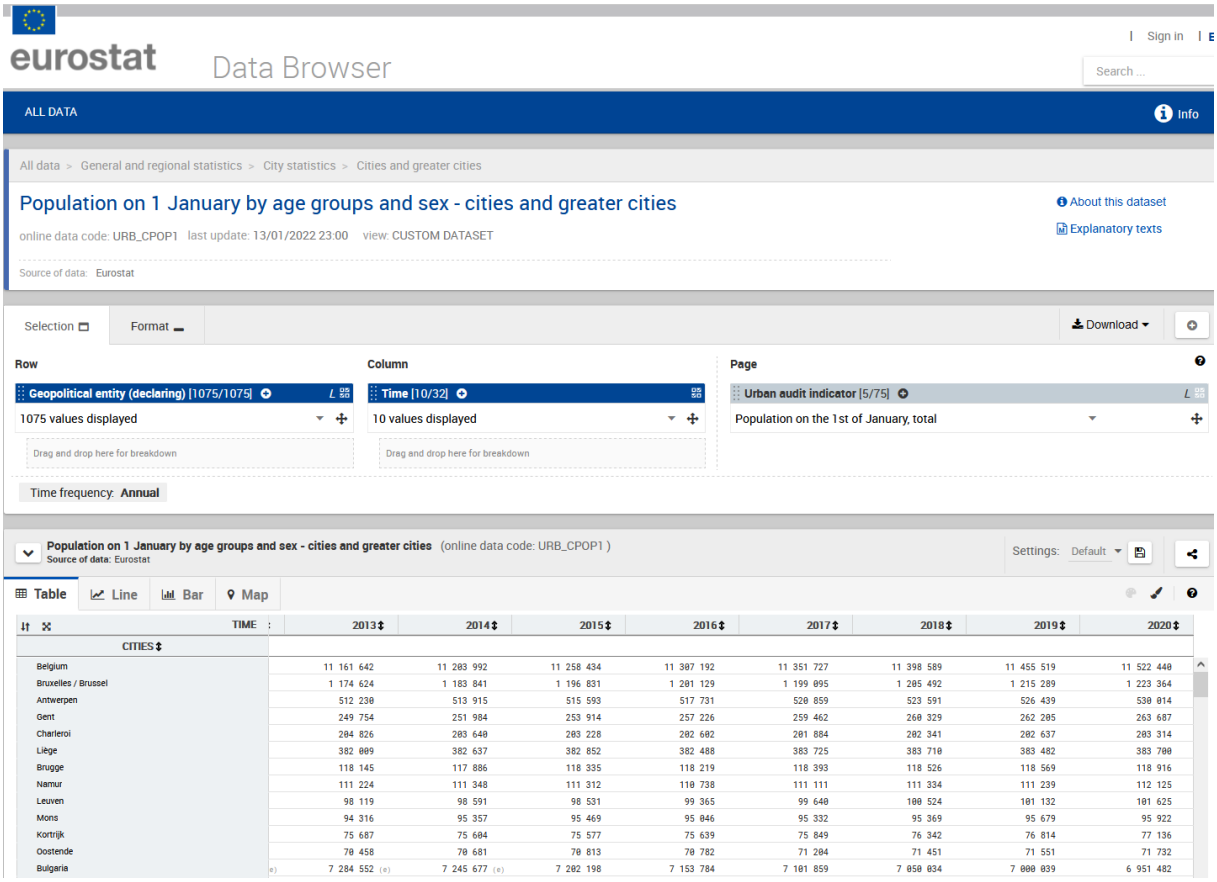
Z obou rozhraní se dostaneme na konkrétní tabulku. Na Obr. 4 je tabulka počtu obyvatel v městech a státech. V záhlaví tabulky je název tabulky, on-line kód (zde např. URB_CPOP1) a datum poslední aktualizace. Kód označení dat je aplikací použit při pojmenování staženého souboru.

V tabulce lze vybírat v záložce *Selection*. Většinou je první vlevo na výběr územní rozsah *Geopolotical entity.* Zde lze vybrat konkrétní stát, město nebo souhrn za 28 zemí EU apod. Počet dostupných a vybraných entit je uveden v záhlaví. V prostřední části *Time* je na výběr časové rozmezí. Ve výchozím nastavení je nabídnuto několik posledních roků. Při požadavku výběru delší historie je třeba zadat a vybrat požadované období. Pro různé statistiky se může lišit rozsah historie a detail údajů (za celé roky, čtvrtletí). V poslední sekci vpravo *Page* jsou na výběr ukazatele, či indikátory. Opět počet ukazatelů závisí na tématu.

Důležité upozornění je, že lze tažením myši zaměnit pozici údaje *Row* a *Column*. Znamená to, že časové údaje jsou jednotlivé řádky a geopolitická entita je ve sloupcích. Možné je i další uspořádání po kliku na tlačítko v tmavě modrém titulku.

Po výběru se okamžitě ukazuje náhled dat v záložce *Table*. Na dalších záložkách lze zobrazit data ve formě grafu *Line* nebo *Bar,* a ve formě mapy – záložka *Map.*

Data se stahují pomocí tlačítka *Download* (vpravo nahoře). Na výběr je více formátů, ve cvičení je používán formát MS Excel .xlsx. Ve staženém souboru bývá několik listů. Na prvních dvou listech *Summary* a *Structure* jsou metadata. Ta obsahují název a kód, přímý link na zdroj dat v databázi Eurostat, datum poslední aktualizace, popis struktury dat, časovou frekvenci dat a jednotku měření. Na třetím listu *Sheet1* a dalších listech *Sheetx* (záleží podle počtu indikátorů) se nachází vlastní data s několika řádkovou hlavičkou, které obsahují datum a čas stažení dat. Pro načtení do software Orange k dalšímu zpracování, doporučuji data zkopírovat na další list a odstranit několik řádků hlavičky, prázdné řádky, vysvětlivky pod tabulkou, tak aby list obsahovat pouze data.



Obr. 4 Výběr parametrů v tabulce Population

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Data extracted on 19/01/2022 11:43:56 from [ESTAT]												
2	Dataset:	Population on 1 January by age groups and sex - cities and greater cities [URB_CPOP1__custom_1933382]											
3	Last updated:	13/01/2022 23:00											
4													
5	Time frequency		Annual										
6	Urban audit indicator		Population on the 1st of January, total										
7													
8	TIME	2011		2012		2013		2014		2015		2016	
9	CITIES (Labels)												
10	Belgium	11 000 638		11 094 850		11 161 642		11 203 992		11 258 434		11 307 192	
11	Bruxelles / Brussel	1 136 778		1 159 448		1 174 624		1 183 841		1 196 831		1 201 129	
12	Antwerpen	498 473		507 368		512 230		513 915		515 593		517 731	
13	Gent	248 358		249 205		249 754		251 984		253 914		257 226	
14	Charleroi	204 150		204 762		204 826		203 640		203 228		202 602	
15	Liège	377 263		379 978		382 009		382 637		382 852		382 488	

Obr. 5 List Sheet1 se staženými daty

V datovém náhledu i stažených datech se může objevit místo číselných údajů **dvojtečka**, což znamená, že data nejsou dostupná. Před další zpracováním je nutná úvaha, jak naložit s chybějícími údaji, záznamy a zda nezkrátit časové rozmezí, vypustiti některé státy, kdy často chybí nejaktuálnější data u některých států.

Další některé příznaky:

Special value

:	not available

Available flags:

b	break in time series
c	confidential data
d	definition differs (see metadata)
e	estimated
f	forecast
r	revised
p	provisional data
U	low reliabilty

Případné sloupce s příznaky je třeba před zpracováním z dat odstranit. Nicméně při intepretaci výsledků je třeba mít tyto informace o datech na paměti a zohlednit je.

> Další informace o databázi Eurostat a dalších datových zdrojích užitečných z pohledu prostorových dat EU lze nalézt v kapitole knihy Spationomy (Pászto et al. 2020).

Úkoly

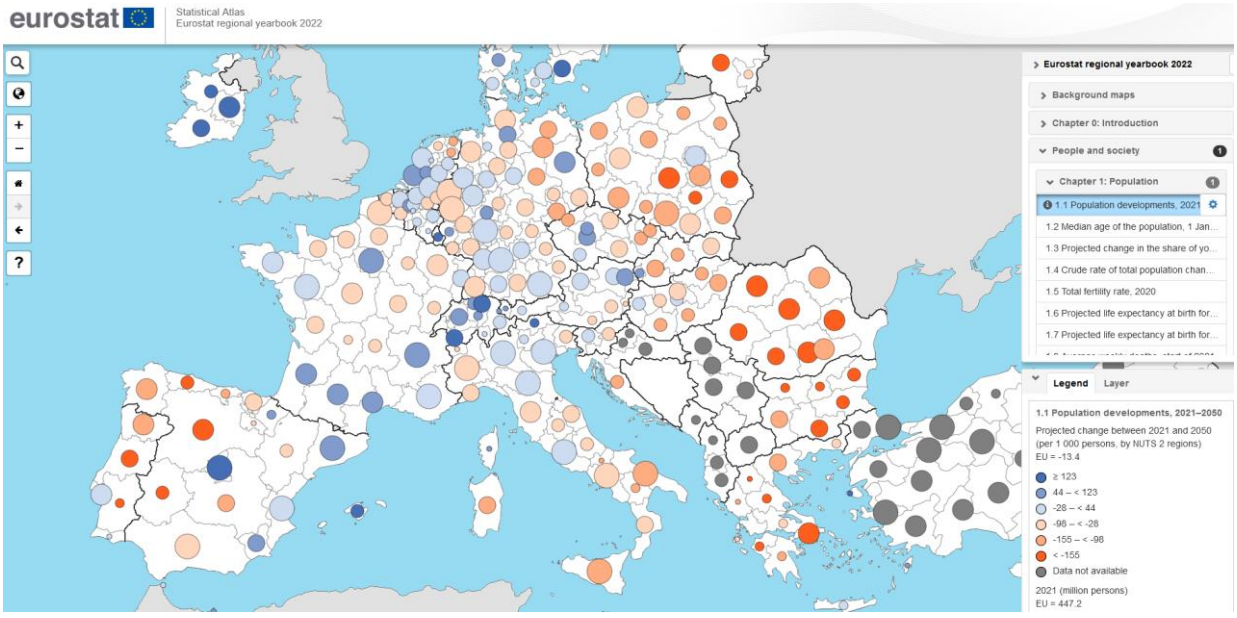
1. Prohlédněte několik témat z databáze Eurostat.
2. Vyzkoušejte nastavení různých parametrů výběru *Geopolitical entity* a *Time*.
3. Stáhněte několik tabulek z různých oblastí témat a prozkoumejte jejich obsah a upravte pro následné zpracování v Orange.
4. Pro konkrétní data rozhodněte úpravy pro chybějící data.
5. Stáhněte si z Eurostatu nejnovější statistickou ročenku **Eurostat regional yearbook** ze sekce Products Statistical Books. *(Pozn. Ve zdrojových datech je ročenka pro rok 2020.)*
6. Dohledejte si odpovídající kapitolu k Vašim datům a nastudujte si charakteristiku příslušného tématu popsanou Eurostatem.

2.1 Statistický atlas Eurostat

Eurostat zobrazuje data také formou **statistického atlasu**, což je interaktivní prohlížeč statistických a topografických map. Statistické mapy používají jako základní podkladové mapy hranice regionů NUTS nebo měst Urban Audit. Data statistického atlasu se čerpají z tištěných verzí regionálních ročenek Eurostatu od roku 2013, souboru dat LUCAS o zjišťování využití půdy/pozemkového pokryvu a dále ze šetření CENSUS 2011 (údaje o sčítání lidu, domů a bytů).

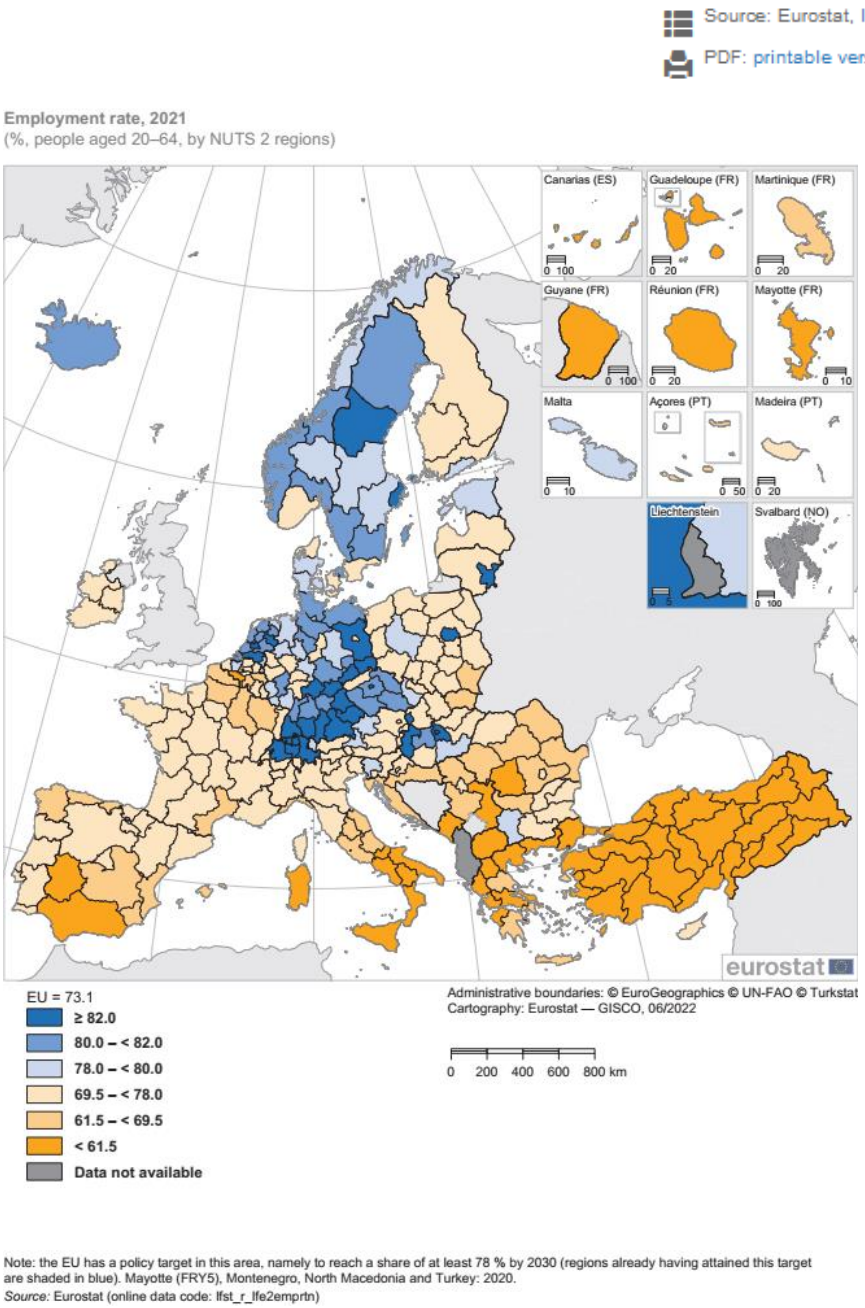
Statistický atlas je dostupný na adrese: *https://ec.europa.eu/eurostat/web/gisco/gisco-activities/statistical-atlas*.

Momentálně je zde jako výchozí dostupný *Eurostat regional yearbook 2022* . Vpravo v okně menu jsou pro výběr dostupné jednotlivé podkladové mapy (NUTS and territorial typologies), dále data z jednotlivých ročenek a v nich detailní kapitoly a oblasti statistických dat.



Obr. 6 Statistický atlas - vývoj počtu obyvatel 2021-2050 v regionech EU

Statistickou mapu lze stáhnout ze Statistického atlasu v PDF formátu. Odkaz je vždy dole v okně menu, kde je uveden i kód zdrojových dat, který funguje jako link na zdrojová data do databáze Eurostat.



Obr. 7 Výstup ze statistického atlasu ve formě automaticky generované PDF mapy, zde zaměstnanost v roce 2021

Úkoly

1. Prohlédněte několik témat z aktuální ročenky s přepnutím různé úrovně regionálního členění (national level, Level 1-3) .
2. Vygenerujte PDF mapy pro jedno vybrané téma z různých roků.

3 COPERNICUS LAND MONITORING SERVICE - URBAN ATLAS

Copernicus Land Monitoring Servis (CLMS) (Agentura pro životní prostředí) EU poskytuje několik datových sad.

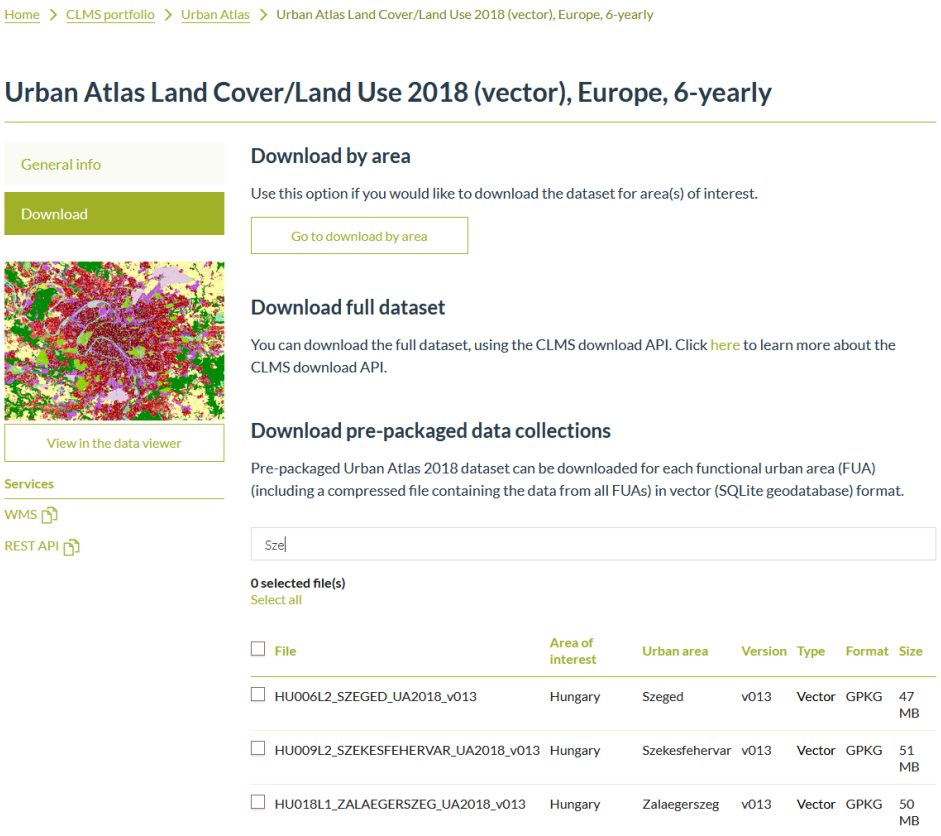
Databáze Urban Atlas je veřejně dostupná na adrese https://land.copernicus.eu/local/urban-atlas.

Urban Atlas obsahuje přibližně 800 samostatných oblastí FUA (Functional Urban Area) pojmenovaných podle největšího města. Nejsou to tedy jen města, ale celé obsáhlé aglomerace s několika sídly. Dostupné jsou tři časové sady z roku 2006, 2012 a 2018. Data z datasetu 2012 a 2018 mají stejnou strukturu členění landuse. Data z roku 2006 mají menší počet kategorií landuse.

Obr. 8 Portál Urban Atlas s možností Download pro dataset z určitého roku

Obr. 9 Mapový náhled dat Urban Atlas

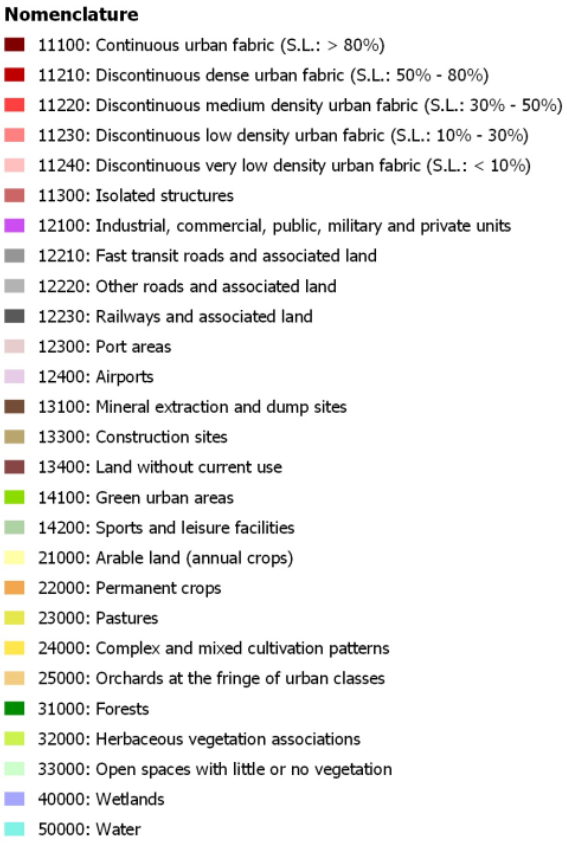
Data jsou ke stažení po přihlášení, je tedy nutná bezplatná registrace do portálu. Funguje jednotné přihlašování formou společného *EU login* pro více agend. Na stránce *Download* je možné vybrat požadovaná FUA ke stažení.



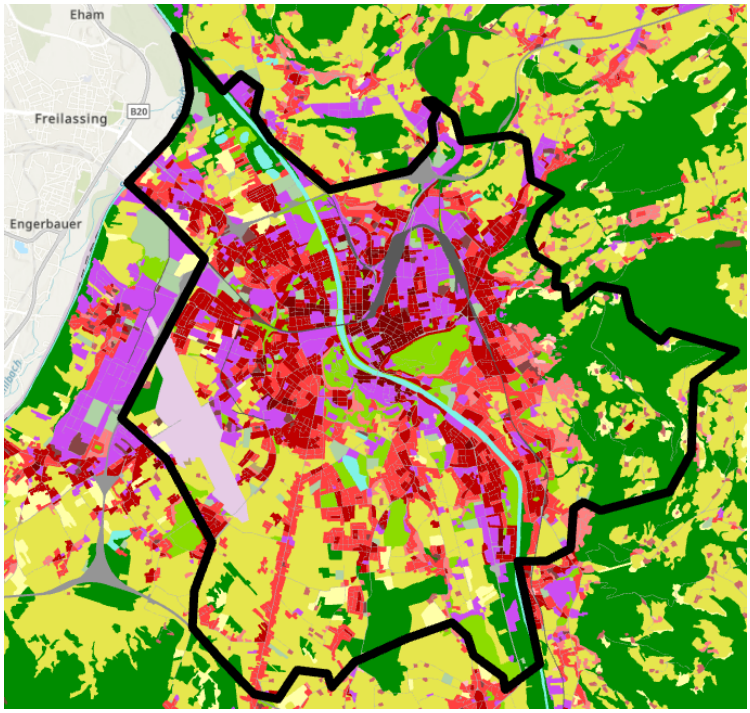
Obr. 10 Vyhledání a stažení dat Urban Atlas

Pro každou oblast je k dispozici jeden zip archiv, který obsahuje: (1) vektorová data ve formátu OGC GeoPackage SQLite (ETRS89-LAEA, EPSG:3035); (2) PDF dokument s mapou oblasti ve vysokém rozlišení; (3) PDF dokument s dodací zprávou; (4) soubory se symbologií ve formátech .lyr, .qml a .sld; a (5) dokument xml s metadaty.

Data obsahují 3 polygonové vrstvy. Kromě polygonů vlastního landuse je to ještě vrtsva *Boundary* a *Core*, které vymezuje oblast centrálního města.



Obr. 11 Kategorie landuse a jejich číselné kódy



Obr. 12 FUA Salzburg s vymezením jádra Core

Úkoly

1. Zaregistrujte se a stáhněte si jedno FUA.
2. Projděte si dokument *Delivery Report* a seznamte se s nomenklaturou kategorií landuse.
3. Nahrajte prostorová data do ArcGIS Pro nebo QGIS a nastavte legendu podle dodaného souboru symbologie.
4. Oříznete z FUA pouze centrum města pomocí vrstvy Core.
5. Zjistěte, které kategorie z landuse chybí (často airport, wetlands apod.).

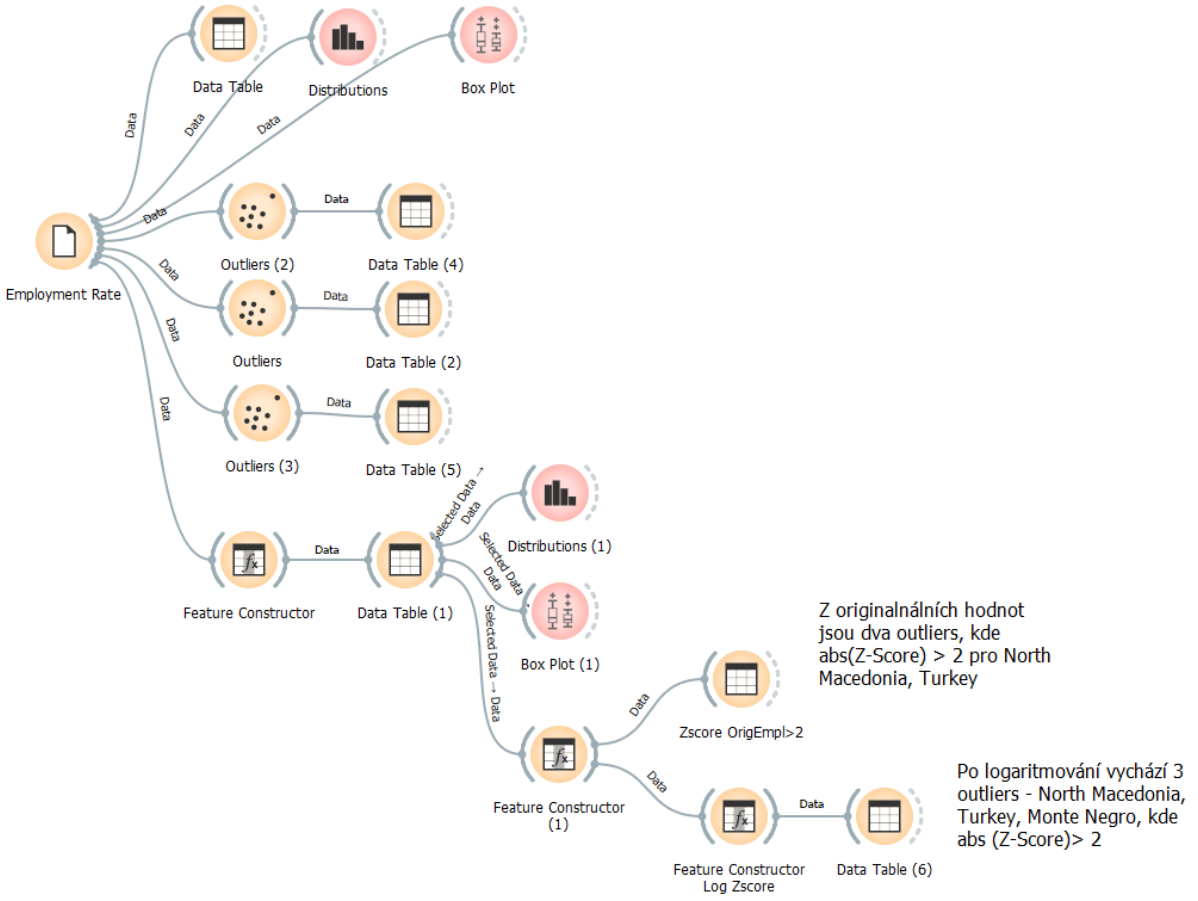
4 ZJIŠTĚNÍ ODLEHLÝCH HODNOT ZAMĚSTNANOSTI STÁTŮ EU

Databáze Eurostat uvádí podíl zaměstnanosti obyvatel v jednotlivých zemích podle pohlaví a celkem. Kódové označení statistiky je **LFSQ_ERGAN**. Úkolem je zjistit, které země lze uvažovat jako odlehlá pozorování, tj. zaměstnanost se výrazně liší od ostatních zemí EU. Odlehlá pozorování jsou zjišťována pro celkovou zaměstnanost. Odlehlá pozorování se budou zjišťovat pomocí různých metod a účelem je porovnat výsledky.

Data *Eurostat_Employment rates.xlsx* (List Q1_2018 obsahuje data za Q1 roku 2018)

Ke stažení https://ec.europa.eu/eurostat/databrowser/view/LFSQ_ERGAN/default/table

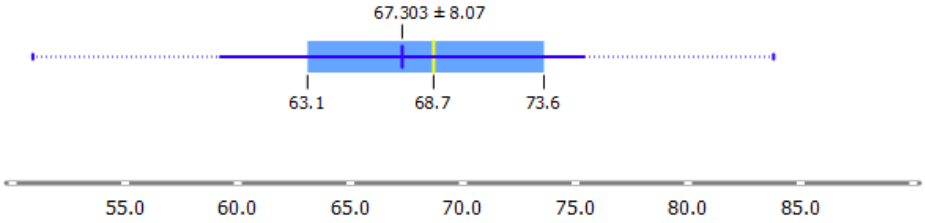
Workflow *Employment.ows*



Obr. 13 Workflow zjišťování odlehlých hodnot

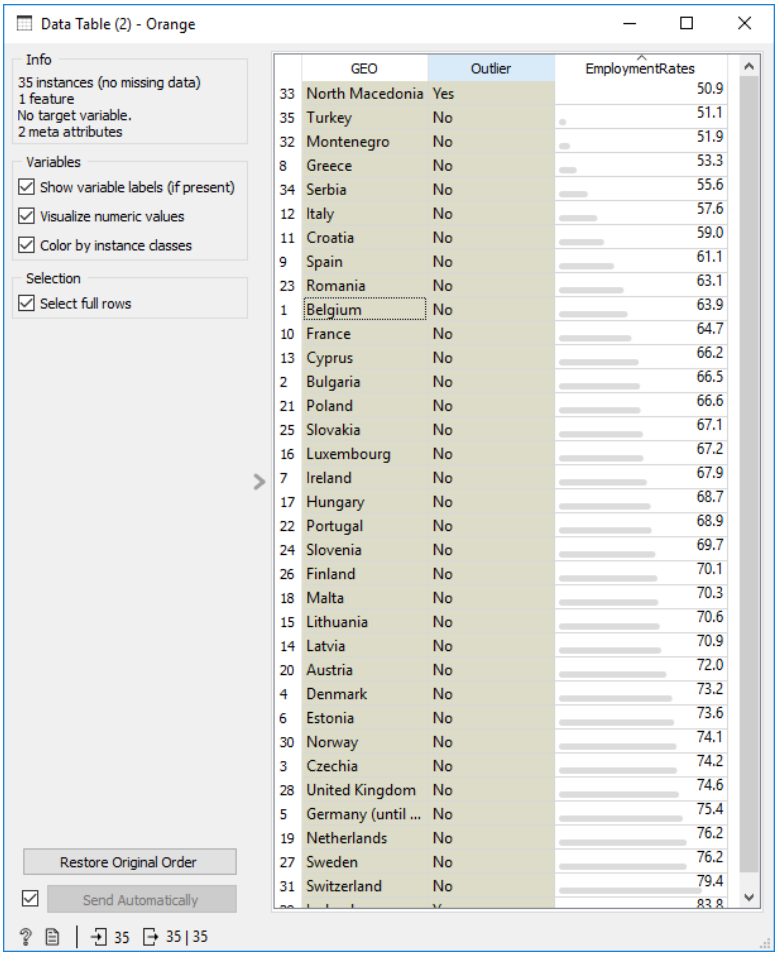
Úkoly

1. Zjistit distribuci hodnot podílu zaměstnanosti, minimální, maximální a průměrnou hodnotu: ……. *(zapište hodnoty, budou třeba pro výpočet Z-score)*
2. Mají hodnoty normální rozdělení (Ano/Ne)?
3. Vykreslete **Box Plot** a zjistěte pomocí něj, zda jsou nějaké hodnoty odlehlé a jakou mají hodnotu.



Obr. 14 Box plot hodnot celkové zaměstanosti v zemích EU

4. Použijte uzel 3x uzel **Outliers** a vyzkoušejte různá nastavení, která poskytuje. Pomocí nově automaticky přidané veličiny *Outliers* v datech (Yes, No) určete, které státy mají výrazně odlišnou zaměstnanost od ostatních států Evropy. Porovnejte s výsledkem z BoxPlotu.



	GEO	Outlier	EmploymentRates
33	North Macedonia	Yes	50.9
35	Turkey	No	51.1
32	Montenegro	No	51.9
8	Greece	No	53.3
34	Serbia	No	55.6
12	Italy	No	57.6
11	Croatia	No	59.0
9	Spain	No	61.1
23	Romania	No	63.1
1	Belgium	No	63.9
10	France	No	64.7
13	Cyprus	No	66.2
2	Bulgaria	No	66.5
21	Poland	No	66.6
25	Slovakia	No	67.1
16	Luxembourg	No	67.2
7	Ireland	No	67.9
17	Hungary	No	68.7
22	Portugal	No	68.9
24	Slovenia	No	69.7
26	Finland	No	70.1
18	Malta	No	70.3
15	Lithuania	No	70.6
14	Latvia	No	70.9
20	Austria	No	72.0
4	Denmark	No	73.2
6	Estonia	No	73.6
30	Norway	No	74.1
3	Czechia	No	74.2
28	United Kingdom	No	74.6
5	Germany (until ...	No	75.4
19	Netherlands	No	76.2
27	Sweden	No	76.2
31	Switzerland	No	79.4

Obr. 15 Výsledek zpracování uzlu Outliers

5. Zjistěte odlehlé hodnoty pomocí spočítání **Z-score**. Z-score nám dobře poslouží pro určení odlehlosti v jednom ukazateli (zde Total). Pro výpočet použijte postupně uzly **Feature Constructor**.
6. Protože rozdělení hodnot zaměstnanosti nemá normální rozdělení, tak původní hodnoty logaritmujte (uzel Feature Constructor).
7. Z-score každé hodnoty Total v konkrétní zemi se počítá jako rozdíl hodnoty X a průměru μ, podělené směrodatnou odchylkou σ. Spočítejte Z-score pro originální hodnoty i logaritmované hodnoty.

$$Z = \frac{X - \mu}{\sigma}$$

8. Jako odlehlé hodnoty určíme ty země, které mají Z-score v absolutní hodnotě větší než číslo 2 abs(Z-score)>2.
9. Vyhodnoťte, zda Z-score pro originální a logaritmované hodnoty indikuje stejné odlehlé hodnoty jako uzly Outliers. Z-score určuje jako odlehlé hodnoty jen země s nízkou zaměstnaností (North Macedonia, Turkey, Montenegro), nikoliv s vysokou zaměstnaností (Iceland). Souhlasí Vaše zjištění?
10. Vyzkoušejte určit odlehlá pozorování pomocí uzlu Outliers, pokud berete v úvahu všechny tři údaje zaměstnanosti (male, female a Total). Ve které zemi je nejmenší a největší zaměstnanost žen? Kde je výrazná disproporce zaměstnanosti žen a mužů?

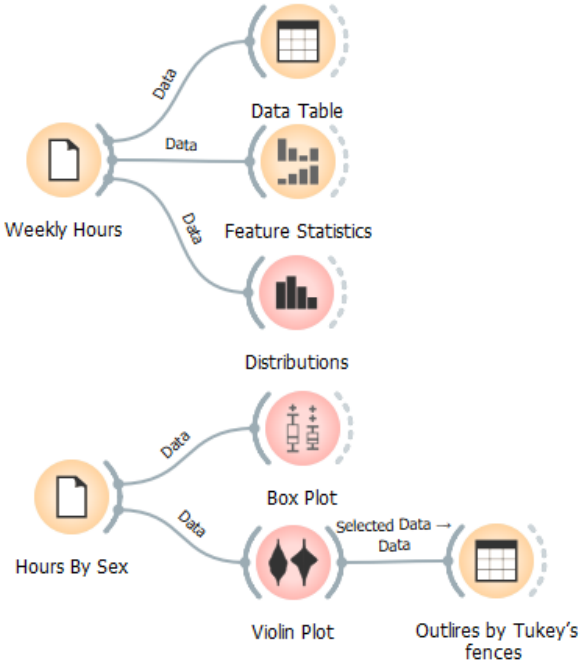
5 POČET PRACOVNÍCH HODIN V TÝDNU PODLE POHLAVÍ VE STÁTECH EU

Z hlediska zaměstnanosti Eurostat eviduje průměrný počet odpracovaných hodin v týdnu v hlavním zaměstnání podle věku, pohlaví a ekonomické aktivity pracujících. Kódové označení statistiky je **LFSA_EWHUN2**. Pro každý stát jsou dostupné ve věkové kategorii 15 až 64 roků celkem tři údaje počtu hodin: pro ženy, pro muže a celkový průměrný počet hodin. Počtem odpracovaných hodin v týdnu se státy EU od sebe podstatně liší. Liší se také průměrným počet hodin mezi skupinou mužů a žen. V tomto příkladu se seznámíte se zajímavým typem grafu – **houslovým (violin) grafem**.

Data *lfsa_ewhun_WorkHours.xlsx* (List Weekly Hours a list Hours By Sex)

Ke stažení
https://ec.europa.eu/eurostat/databrowser/view/LFSA_EWHUN2/default/table?lang=en&category=labour.employ.lfsa.lfsa_wrktime

Workflow *WorkingHours_ViolinPlots.ows*

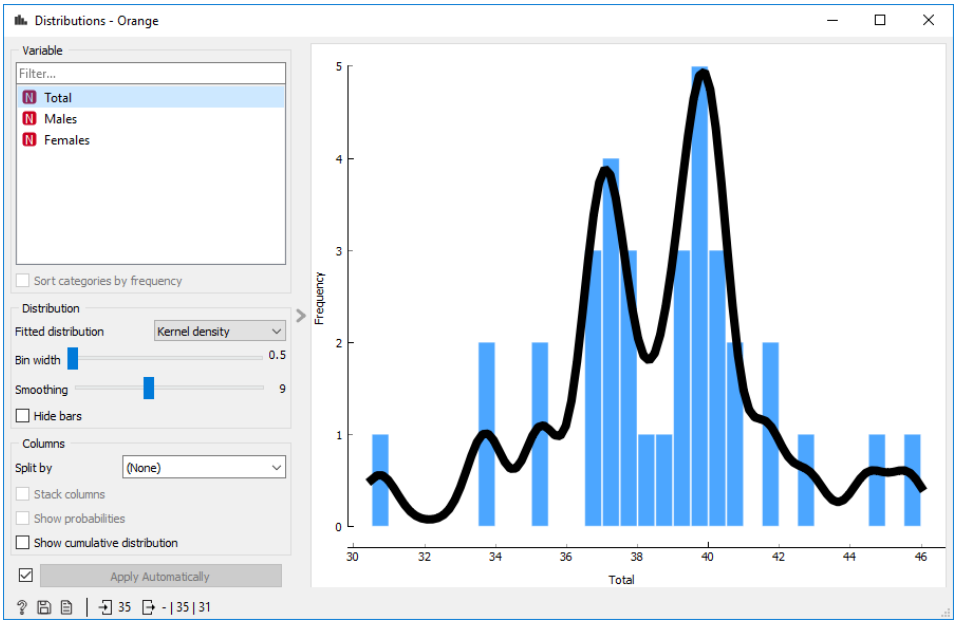


Obr. 16 Workflow pro zpracování průměrného počtu hodin

Úkoly

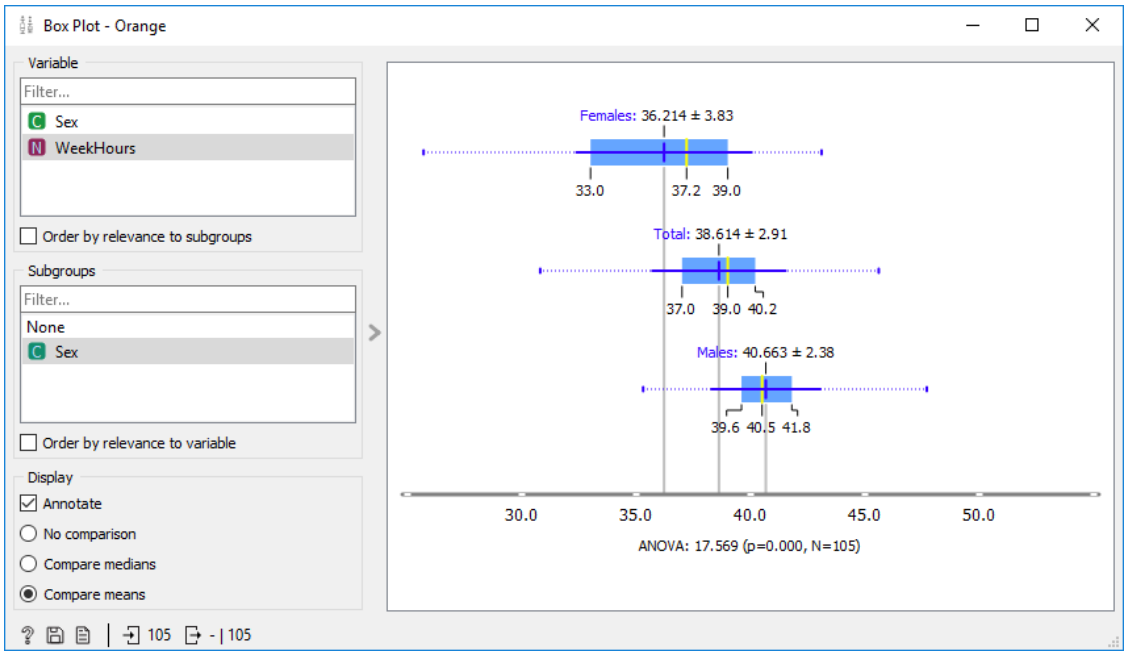
1. Zjistit distribuci, maxima a minima průměrného počtu hodin. Zdrojem je list *Weekly Hours*, kde jsou pro každý stát tři sloupce s uvedením počtu hodin pro ženy, muže a celkem (Total).
2. Ve které zemi je nejnižší, nejvyšší počet hodin podle různých kategorií?
   Vypište: ……

3. Distribuce celkového počtu má dva vrcholy z důvodu dvou vrcholů v histogramu pro ženy a muže. Fitujte rozdělení distribucí *Kernel density*, protože ta je potom použita při výpočtu a zobrazení u houslového grafu.



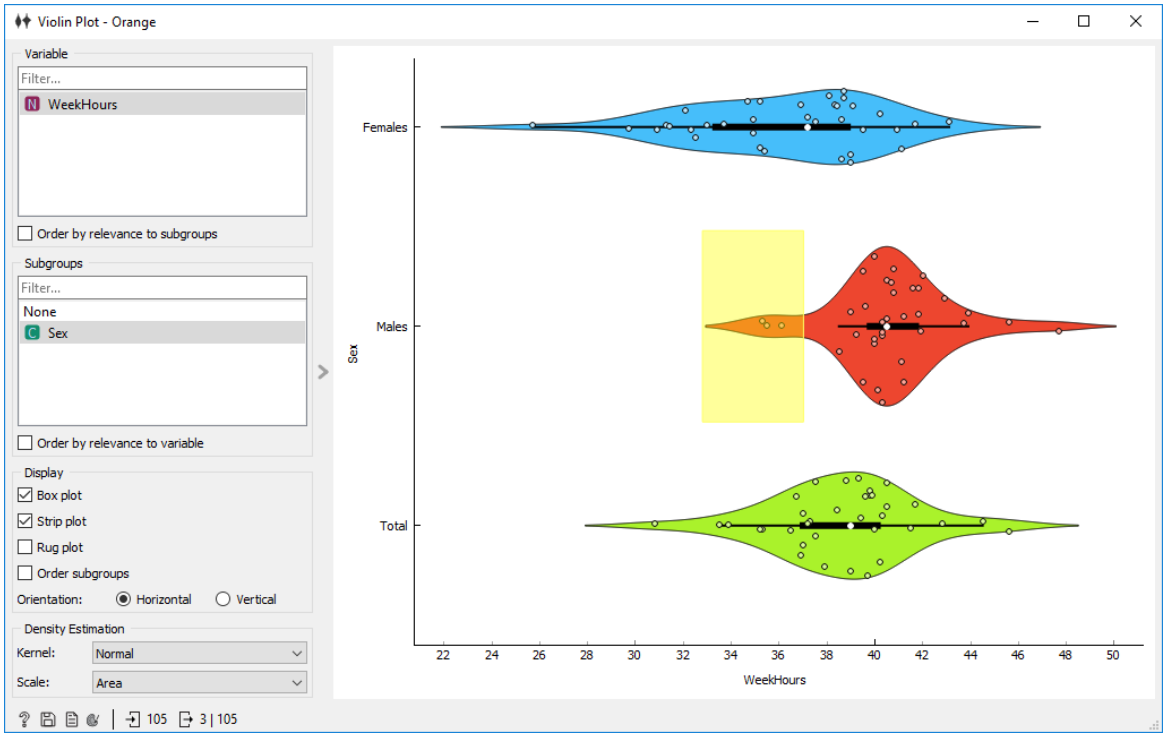
Obr. 17 Distribuce počtu hodin

4. Data na listu *Hours By Sex* obsahují upravená data, kde je jen jeden sloupec s počtem hodin, dále kategoriie *Sex* a opakovaně názvy států. Toto uspořádání dat umožní zobrazit data v boxplotu a houslovém grafu zároveň podle kategorie *Sex*.



Obr. 18 Box plot počtu pracovních hodin podle kategorie pohlaví

5. Komentujte na základě boxplotu odlišnosti mezi počtem hodin mužů, žen a identifikujte extrémní hodnoty.
6. Vykreslete houslový graf pomocí uzlu **Violin Plot**, opět nastavte samostatné grafy podle pohlaví Sex.
7. Zaškrtněte volbu *Strip plot*, kdy se vykreslí body vstupních hodnot. Lze vybrat různé hodnoty Kernelu: *Normal, Epanechnikov* a *Linear*.



Obr. 19 Violin graf počtu pracovních hodin v týdnu podle pohlaví

8. **Houslový graf** je specifický graf, který aplikuje kernel density estimation (jádrový odhad hustoty) na bodová data. Znamená to, že kde je graf širší, tak se nachází více bodů v okolí. Naopak, kde je graf užší, tak je bodů v okolí méně. Nejedná se tedy o historgramy!

9. Porovnejte šířku grafu jednotlivých kategorií žen a mužů, který je užší a který je širší? ……

10. Který je protáhlejší? ……

11. Houslový zobrazuje některé charakteristiky jako boxplot (zatržítku Box plot). Bílým bodem uprostřed je zobrazen medián, silná černá čára uprostřed zobrazuje mezikvartilové rozpětí jako krabice u boxplotu. Tenké černé čáry zobrazují **Tukey's fences**. Oblasti za touto čárou vymezují odlehlé hodnoty. Tukey's fences lze spočítat podle vzorce $[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)]$,

    kde *Q1* a *Q3* je první a třetí kvartil a konstanta *k* při hodnotě k = 1,5 indikuje odlehlé hodnoty. Při hodnotě k = 3 identifikuje data, která jsou zcela mimo rozsah. V Orange je vykresleno pro k=1,5.

12. Vyznačte tažením myši v houslovém grafu některou z oblastí nad/pod tekou čárou. Je zobrazena žlutým obdélníkem. Jsou tak vybrány body, která lze následně zobrazit v uzlu *Data Table* ve workflow a určit tak odlehlé hodnoty. Pozor, takto lze určit pouze odlehlé hodnoty v rámci jednoho atributu. Není toto vyšetřování vhodné pro multidimenzionální data.

> *Vysvětlení a popis použití houslového grafu lze nalézt na Orange blogu v příspěvku Box Plot Alternative: Violin Plot* (Pretnar 2021).

6 SHLUKY STÁTŮ PODLE POČTU PRACOVNÍCH HODIN

V tomto příkladu budou využita stejná data jako v příkladu předchozím. Budeme hledat shluky evropských států, které se podobají počtem pracovních hodin mužů a žen. Bude použita shlukovací metoda k-Means.

Data *lfsa_ewhun_WorkHours.xlsx* (list Weekly Hours)

Ke stažení
https://ec.europa.eu/eurostat/databrowser/view/LFSA_EWHUN2/default/table?lang=en&category=labour.employ.lfsa.lfsa_wrktime

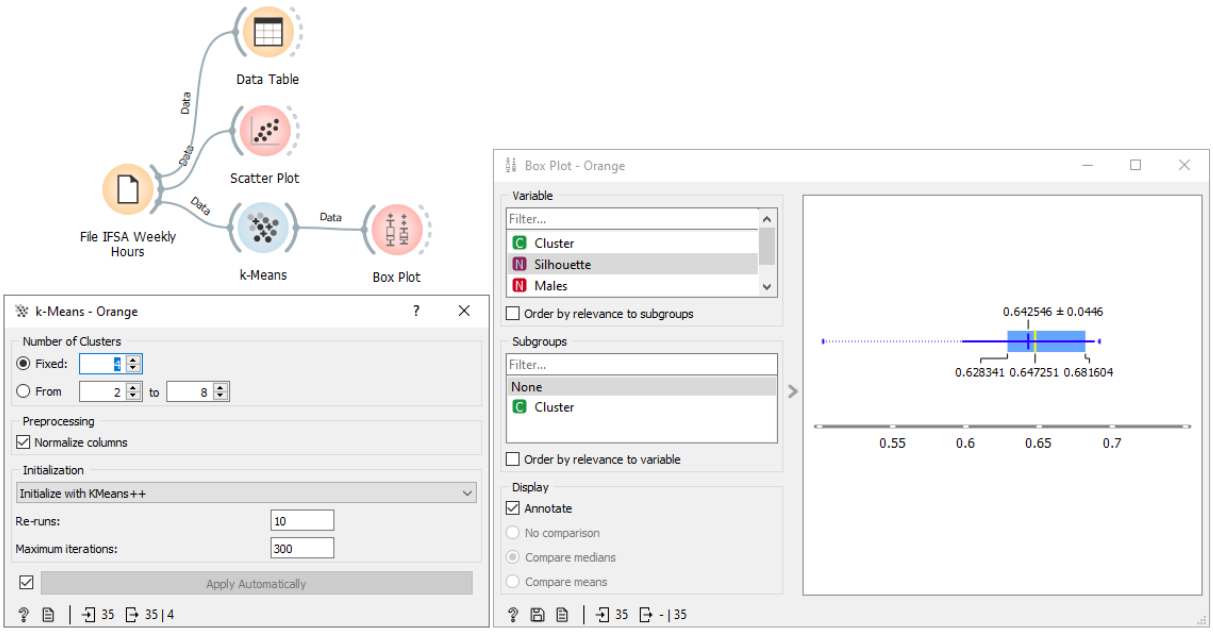
Workflow *WorkingHours_Clusters.ows*

Inspirace How to choose k for k-Means? (*https://www.youtube.com/watch?v=TKYAeY3lUJc)*

Kvalita shlukování lze vyjádřit pomocí indexu siluety. Ta je automaticky počítána pro každý objekt ve výstupních datech. Při hledání nejlepšího počtu shluků – k v metodě k-Means, můžeme pomocí boxplotu vyšetřovat průměrnou hodnotu siluety pro každou hodnotu k.

Úkoly

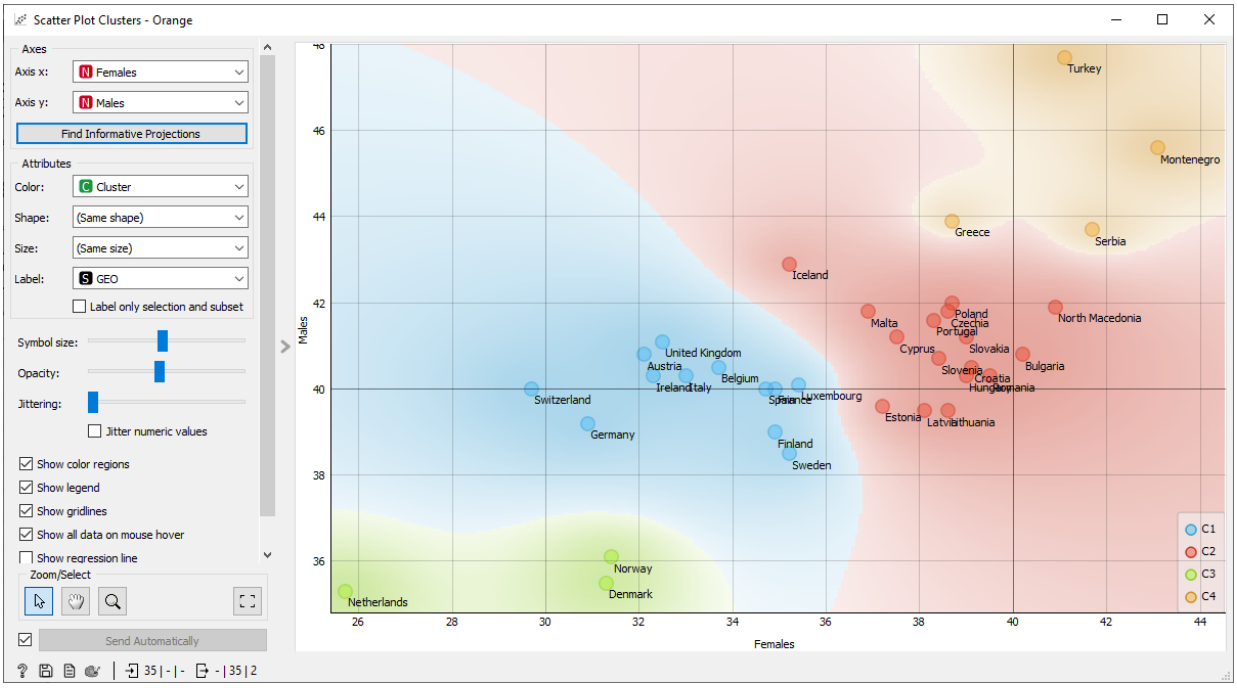
1. Zobrazte zároveň dialog uzlu *k-Means* a připojte za něj uzel *Box Plot*. Při změně hodnoty v kolonce *Fixed* (počet shluků) se bude měnit boxplot a průměrná hodnota siluety. V případě těchto dat je nejvyššího průměrná hodnota 0,64 (modrá čára) pro 4 shluky (Obr. 20).
2. Vyberte volbu *From … to* v uzlu k-Means, potom se průměrná hodnota siluety objeví v levém okně dialogu. Postupně experimentálně zkontrolujte, že hodnoty zadané pomocí *Fixed* souhlasí s hodnotami průměru zobrazeným v boxplotu. Hodnota indexu průměrné siluety bude postupně růst a potom klesat.



Obr. 20 Experimentování s hodnotou k a vykreslení boxplotu indexu siluety všech objektů pro vybrané k

3. Vykreslete výskytový graf Scatter Plot a zobrazte shluky států (Obr. 21).

4. Shluky interpretujte.

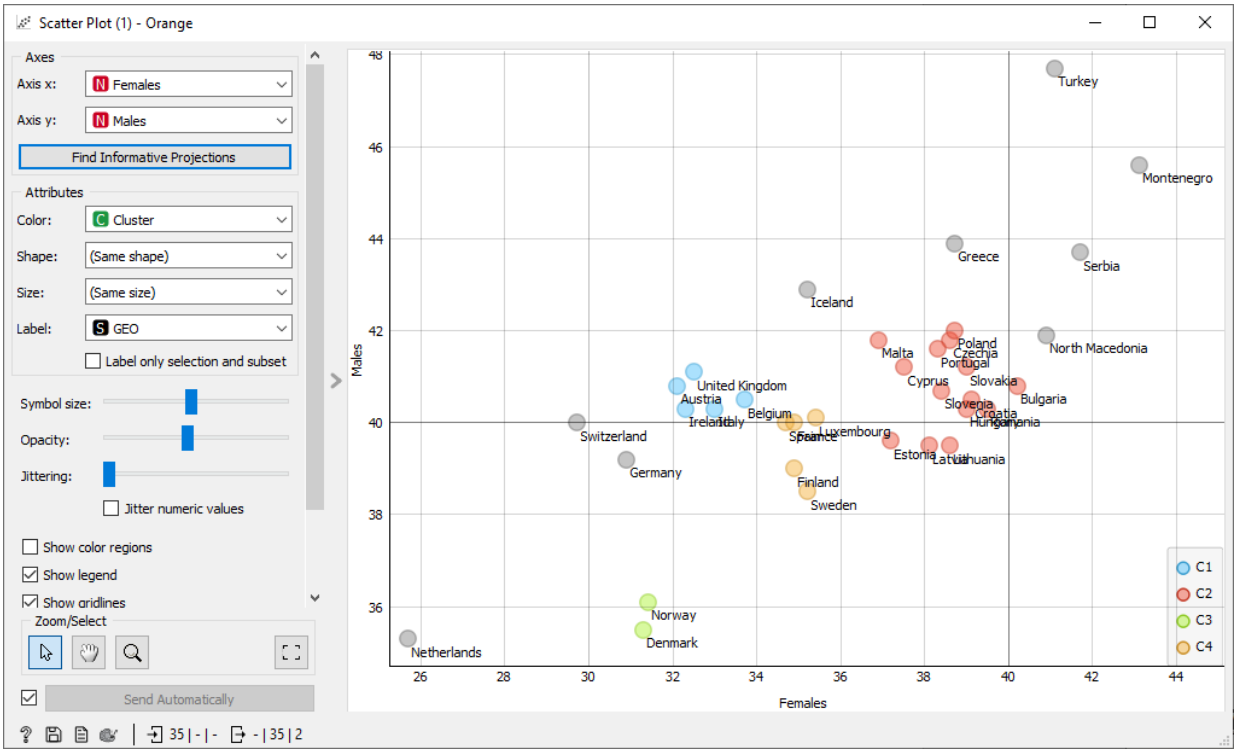


Obr. 21 Shluky států podle počtu pracovních hodin v týdnu pro muže a ženy

Vyzkoušejte i shlukování metodou **DBSCAN** (*Core point neigbors 2, Distance 1*). Výsledkem jsou méně početné 4 shluky, ale i šumové body. Odlehlé státy jsou právě označeny jako šumové body. Ve větším detailu je rozlišen střed – modrý, žlutý a červený shluk. Zajímavý je dvou prvkový shluk Norsko a Dánsko.

Která metoda shlukování, k-Means nebo DBSCAN, dává lepší výsledek?

Porovnejte, který výsledek shlukování se lépe interpretuje.



Obr. 22 Výsledné shluky vyšetřené metodou DBSCAN

7 ZAMĚSTNANOST OSOB V ZEMÍCH EU PODLE EKONOMICKÉ AKTIVITY

Databáze Eurostat sleduje zaměstnanost osob podle oblastí jednotlivých ekonomických aktivit. Kódové označení statistiky v databázi Eurostat je **NAMA_10_A64_E**. Věkové rozmezí zaměstnaných osob je ve věku od 15 do 64 roků.

Je použita klasifikace NACE Rev. 2, která byla přijata v roce 2006 (Eurostat 2022).

Klasifikace NACE má členění na čtyři hierarchické úrovně (European Commission 2008):

Level 1: 21 skupin označené písmeny A až U;

Level 2: 88 skupin označené dvouciferným číselným kódem (01 až 99);

Level 3: 272 skupin označené trojciferným číselným kódem (01.1 to 99.0);

Level 4: 629 tříd označené čtyřciferným číselným kódem (01.11 to 99.00).

Na úrovni 1 se ekonomické aktivity dělí do oblastí dle Tab. 1

Tab. 1 Kódy a název skupiny ekonomické aktivity na úrovni Level 1

Code	Description
A	AGRICULTURE, FORESTRY AND FISHING
B	MINING AND QUARRYING
C	MANUFACTURING
D	ELECTRICITY, GAS, STEAM AND AIR CONDITIONING SUPPLY
E	WATER SUPPLY; SEWERAGE, WASTE MANAGEMENT AND REMEDIATION ACTIVITIES
F	CONSTRUCTION
G	WHOLESALE AND RETAIL TRADE; REPAIR OF MOTOR VEHICLES AND MOTORCYCLES
H	TRANSPORTATION AND STORAGE
I	ACCOMMODATION AND FOOD SERVICE ACTIVITIES
J	INFORMATION AND COMMUNICATION
K	FINANCIAL AND INSURANCE ACTIVITIES
L	REAL ESTATE ACTIVITIES
M	PROFESSIONAL, SCIENTIFIC AND TECHNICAL ACTIVITIES
N	ADMINISTRATIVE AND SUPPORT SERVICE ACTIVITIES
O	PUBLIC ADMINISTRATION AND DEFENCE; COMPULSORY SOCIAL SECURITY
P	EDUCATION
Q	HUMAN HEALTH AND SOCIAL WORK ACTIVITIES
R	ARTS, ENTERTAINMENT AND RECREATION
S	OTHER SERVICE ACTIVITIES
T	ACTIVITIES OF HOUSEHOLDS AS EMPLOYERS; UNDIFFERENTIATED GOODS- AND SERVICES-PRODUCING ACTIVITIES OF HOUSEHOLDS FOR OWN USE
U	ACTIVITIES OF EXTRATERRITORIAL ORGANISATIONS AND BODIES

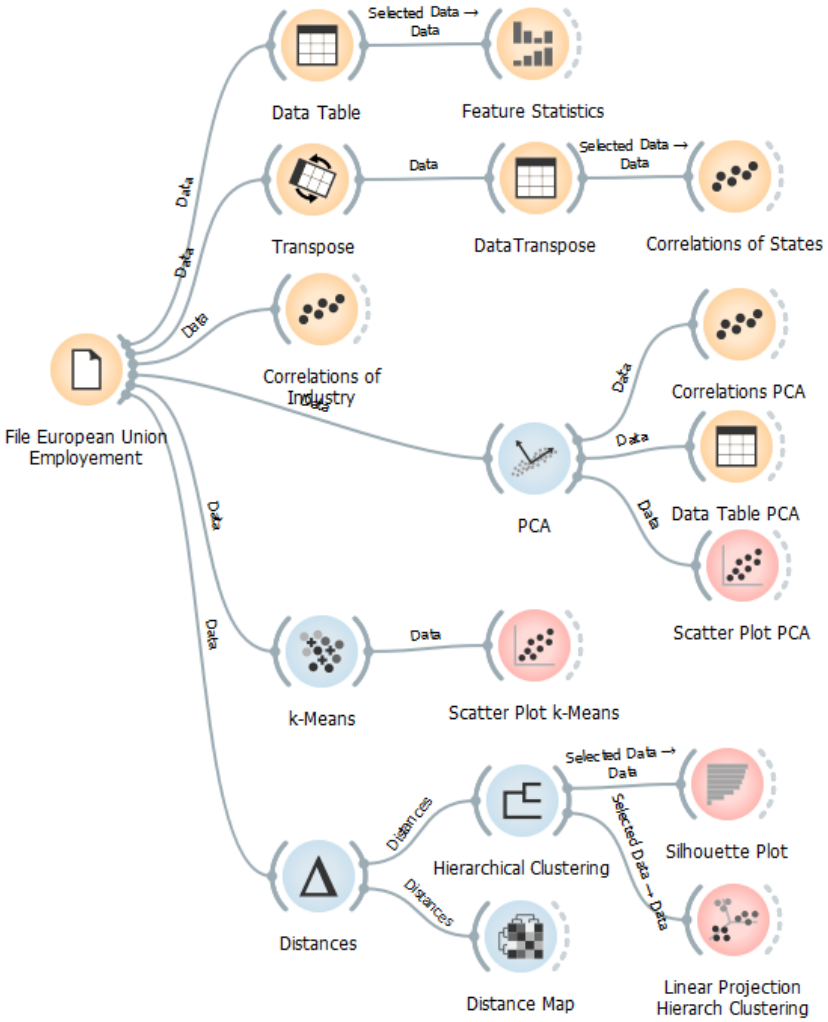
> *Příklad vychází ze článku* (Masopust et al. 2021).

Data *nama_10_a64_e_2018.xlsx*

Ke stažení *https://ec.europa.eu/eurostat/databrowser/view/NAMA_10_A64_E*

Data Eurostat obsahují údaje do úrovně v úrovni Level 1a Level 2. Ve cvičných datech je vybrána pouze úroveň 1 a rok 2018, který je momentálně nejúplnější. Stažená data mají měrnou jednotku tisíce osob nebo lze vybrat odpracované hodiny podle ekonomické aktivity (tis.). Pro zkoumání podobnosti byla data následně přepočítána na procentuální zastoupení jednotlivých aktivit, tak aby státy byly souměřitelné. Nejprve se provede součet všech osob za jednotlivé státy. Následně se spočítá procentuální podíl v konkrétním státě v jednotlivých ekonomických aktivitách. U nových dat je nutný tento přepočet před vlastním zpracováním.

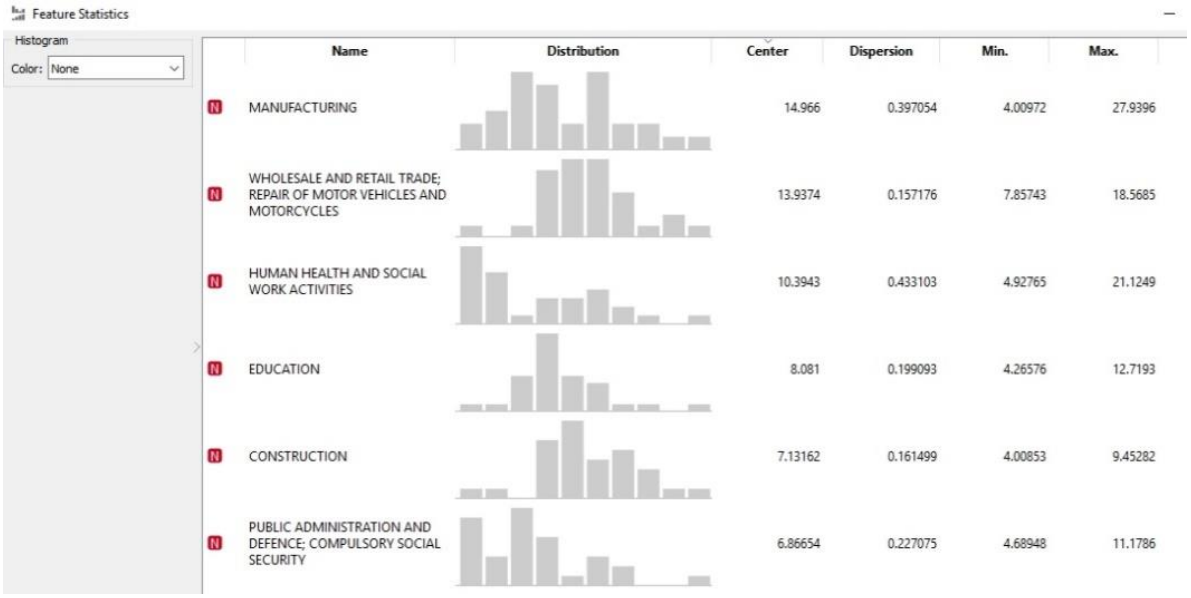
Workflow



Obr. 23 Workflow pro zpracování údajů o pracujících v oblastech ekonomických aktivit

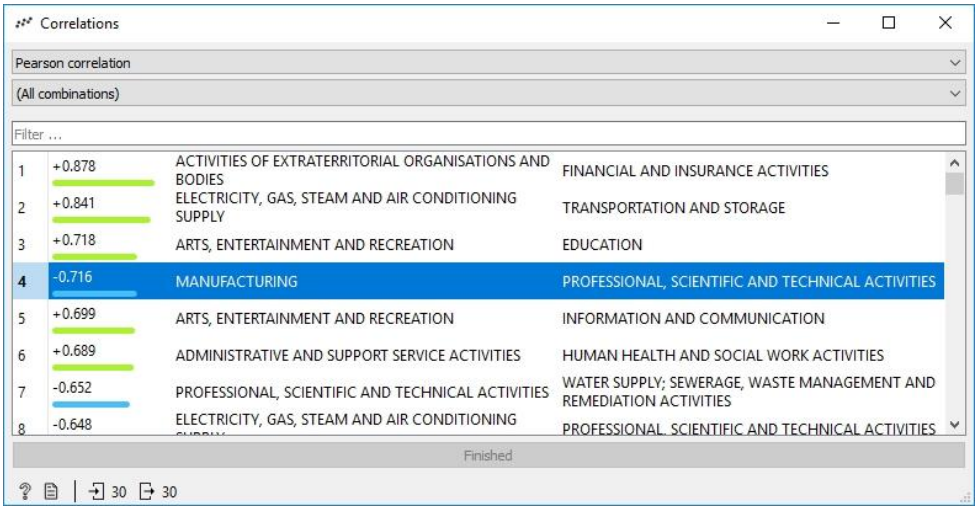
Úkoly

1. Proveďte základní charakteristiku dat pomocí uzlu *Data Table* a *Feature Statistics*. Zjistěte státy, kde je zaměstnáno nejvíce a nejméně pracujících podle ek. aktivit. Která státy a ve kterých aktivitách se pohybují kolem průměru EU? Jednotlivá zjištění průběžně zapisujte.



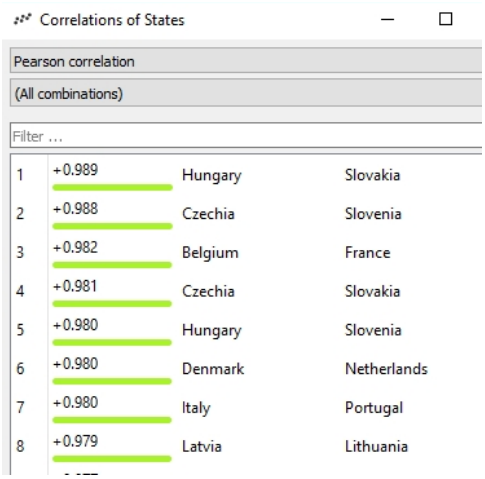
Obr. 24 Feature Statistics podle skupin aktivit

2. Vypočítejte korelace mezi ekonomickými aktivitami pomocí uzlu *Correlations*. Určete nejvyšší kladné korelace a nejnižší záporné korelace.



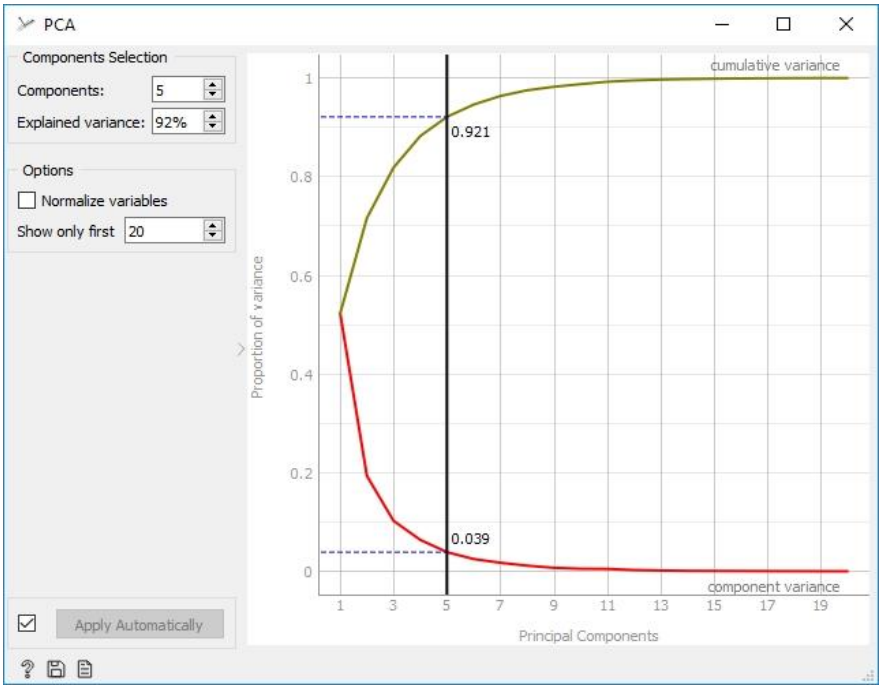
Obr. 25 Korelace skupin aktivit

3. Vypočítejte korelace mezi jednotlivými státy pomocí uzlu *Correlations*. Je nejprve provést transponování vstupních dat pomocí uzlu *Transpose*. Které dvojice států nejvíce korelují, a které korelují nejméně a které mají nejvyšší negativní korelaci?



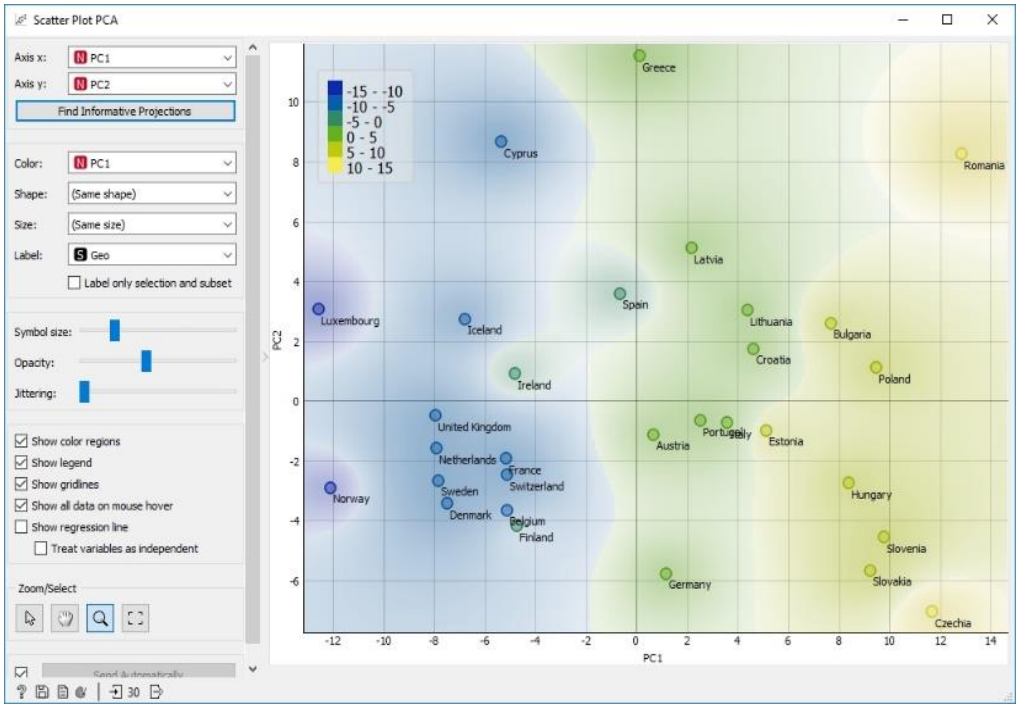
Obr. 26 Korelace států

4. Proveďte analýzu hlavních komponent pomocí uzlu *PCA.* Určete vhodný počet nových komponent a zapište procenta vysvětlované variability. Nalezněte nové pojmenování pro komponenty PC1, PC2.



Obr. 27 Scree graf (graf vlastních čísel) PCA

5. Vykreslete rozptylový graf pomocí uzlu *Scatter Plot* v souřadnicích PC1 a PC2. Vyšetřete shluky podobných států, nalezněte odlehlé státy, které si nejsou výrazně podobné s dalšími státy. Který stát je nejblíže středu 0, 0?

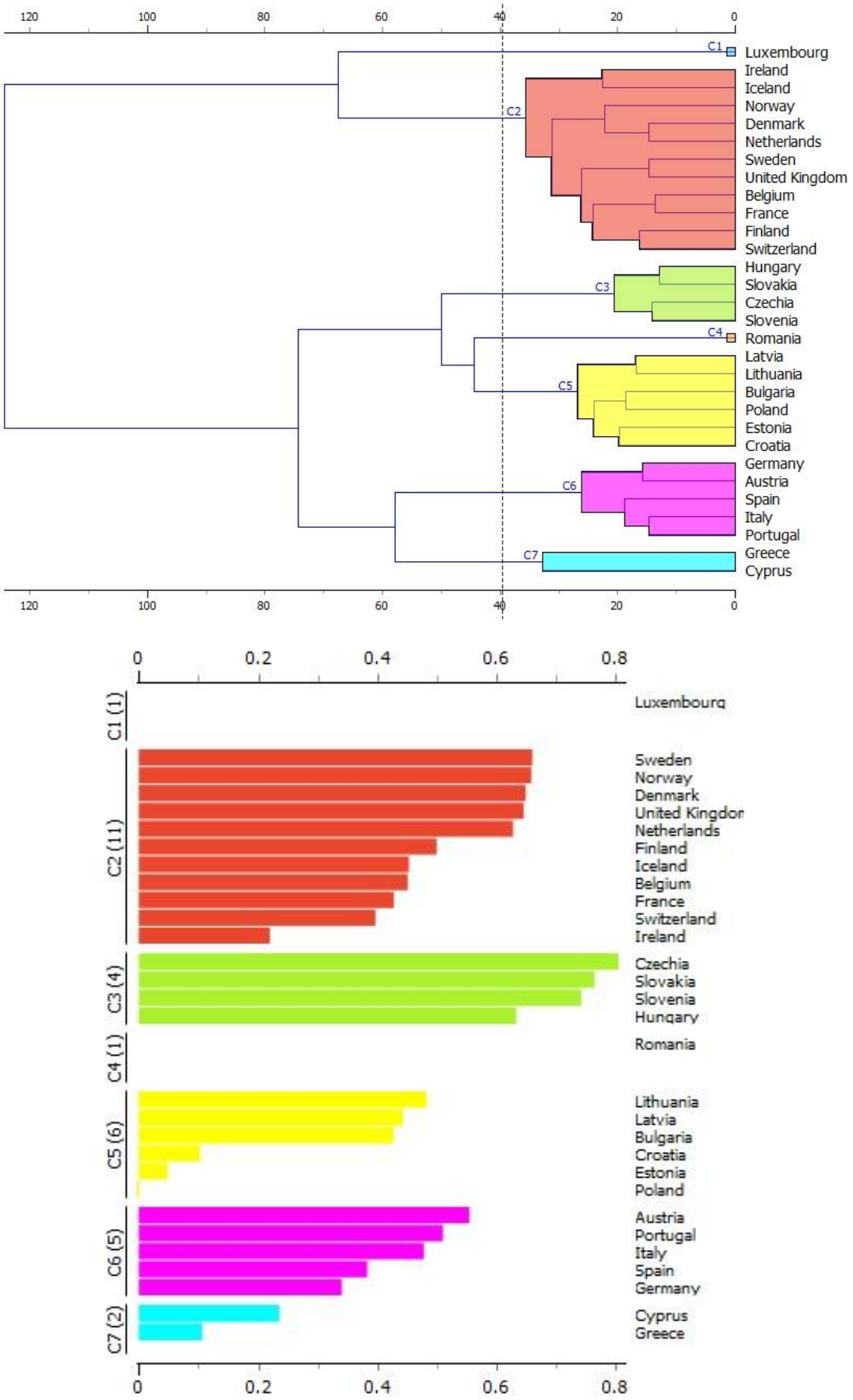


Obr. 28 Scree graf PCA

6. Proveďte shlukování metodou *k-Means*. Podle údaje *Silhoute score* nastavte vhodný počet shluků a experimentujte jak s jiným počtem shluků, tak s různým způsobem inicializace shlukování - *Random initialisation* nebo *Initialisation with KMeans++.* Pro výsledek shlukování vykreslete rozptylový graf s nabídnutou nejlepší dvojicí souřadnic. Vykreslete pro výsledek k-Means graf siluety shluků pomocí uzlu *Silhoutte plot.*

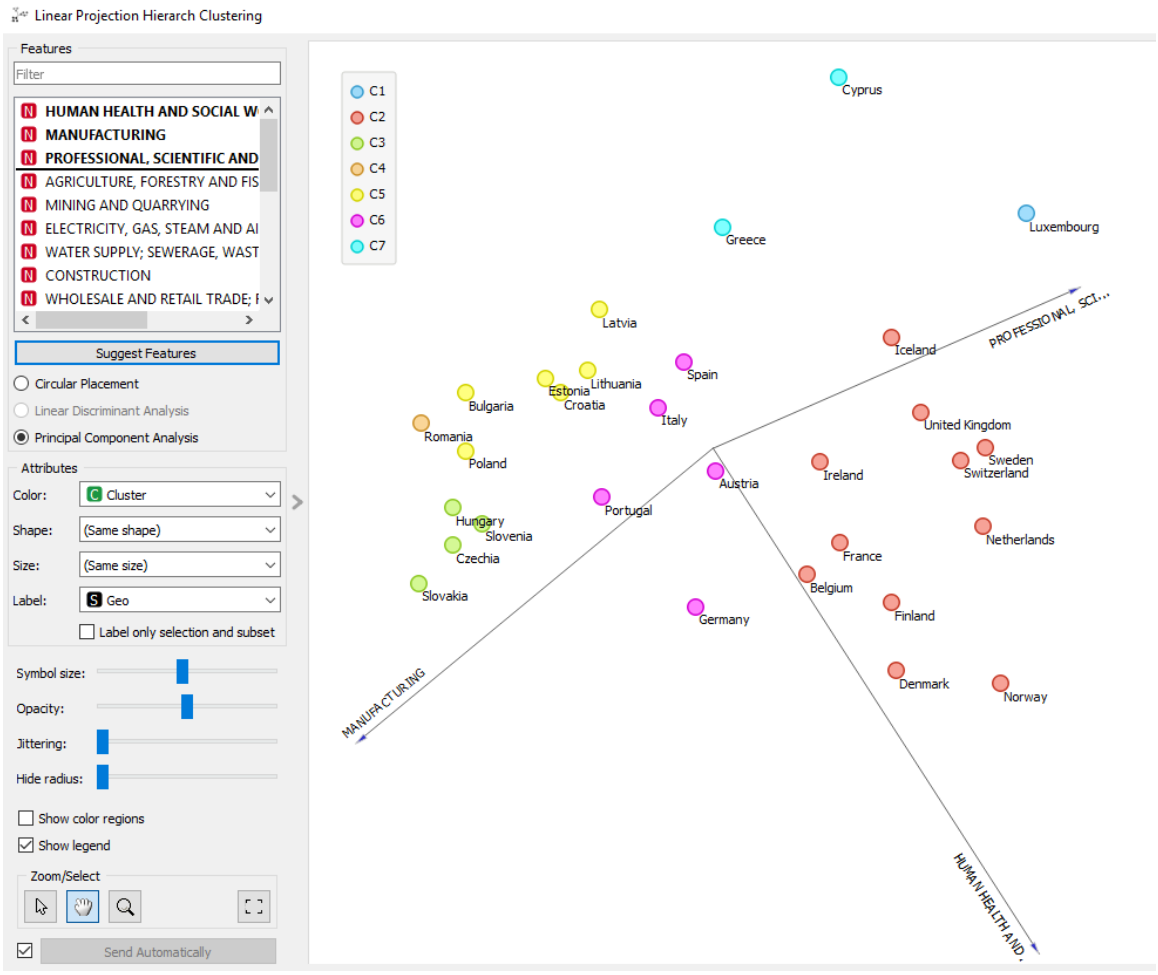
Obr. 29 Rozptylový graf výsledku shlukování metodou k-Means

7. Proveďte hierarchické shlukování. Pro uzel *Distances* nastavte metriku *Cosine* (můžete zkusit i Manhattan nebo Mahalanobise). V dendrogramu uzlu *Hierarchical clustering* nastavte Wardovu metodu výpočtu mezishlukové vzdálenosti, dále vyberte vhodnou úroveň pro vymezení počtu shluků. Kvalitu výsledných shluků posuďte pomocí *Silhoute plot*. Můžete udělat několik stejných větví heirarchického zpracování s různými parametry a následně porovnávat výsledky.



Obr. 30 Hierarchické shlukování a jeho graf siluety

8. Zobrazte výsledek hierarchického shlukování pomocí uzlu *Linear Projection*.



Obr. 31 Zobrazení shluků určeného hierarchickým shlukováním ve třech doporučených osách

Závěrem porovnejte dílčí výsledky. Sledujte, kdy dostáváte obdobné informace, např. stejné skupiny států jsou si podobné podle různých metod. Zjistěte, zda stejné oblasti ekonomických aktivit převažují ve stejné skupině států atd. Zjistěte, zda existuje rozdělení na západní a východní evropské ekonomiky

Volitelné úkoly

1. Celý postup zkuste pro údaje ekonomických aktivit pro data **Level2** (88 údajů pro každý stát).
2. Celý postup zpracování zopakujte pro údaj **odpracovaných hodin** podle druhu ekonomické aktivity.

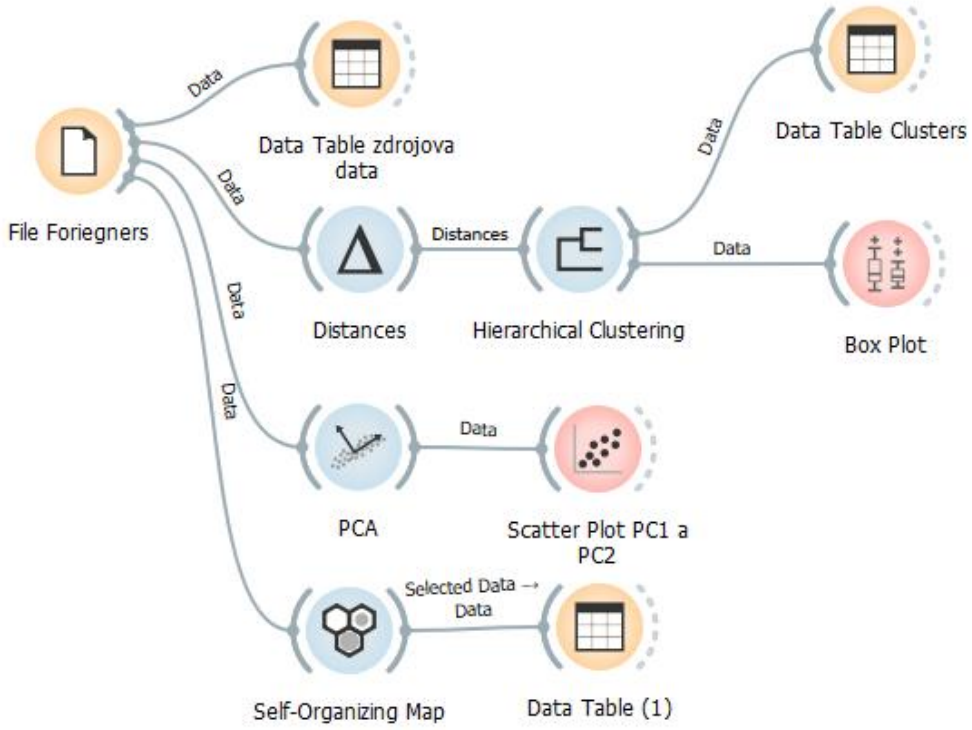
8 INTERPRETACE SHLUKŮ STÁTŮ PODLE POČTU CIZINCŮ

Databáze Eurostat poskytuje statistiku počtu cizinců žijících v jednotlivých zemích EU. Statistika je z roku 2011, má kód CENS_11CTZ_N. Pro každý stát jsou uvedeny počty cizinců podle kontinentu, odkud přišli. Data jsou uvedena podle zdrojové oblasti na: *EU-countries except reporting country (evropské státy bez sledované země)*, *European non-EU countries (Evropské země bez států EU)*, *Afrika, Severní Amerika, Karibská oblast a centrální a Jižní Amerika, Asie, Oceánie* a bez udání zdrojové země (*Unknown*). Navíc jsou v datech kontrolní součty (počet z neevropských zemí a celkový počet). Dva atributy, které jsou kontrolní součty (*Total…*) nevstupují do zpracování – vyšetření shluků, nastavte tedy skip v uzlu File.
Vstupní data počtu cizinců byla pro účely tohoto příkladu přepočítána na relativní zastoupení cizinců vůči celkovému počtu obyvatel každé země (údaj Reporting country).

Inspirace *Explaining Clusters* video https://www.youtube.com/watch?v=3SkjU2eBzNY&t=9s (data set Course Grades) nebo https://www.youtube.com/watch?v=3SkjU2eBzNY

Data *cens_11ctz_Foreigners.xlsx (list Procenta Cizinci)*

Workflow *Foreigners_Explain.ows*



Obr. 32 Workflow pro inspekci počtu cizinců žijících v EU

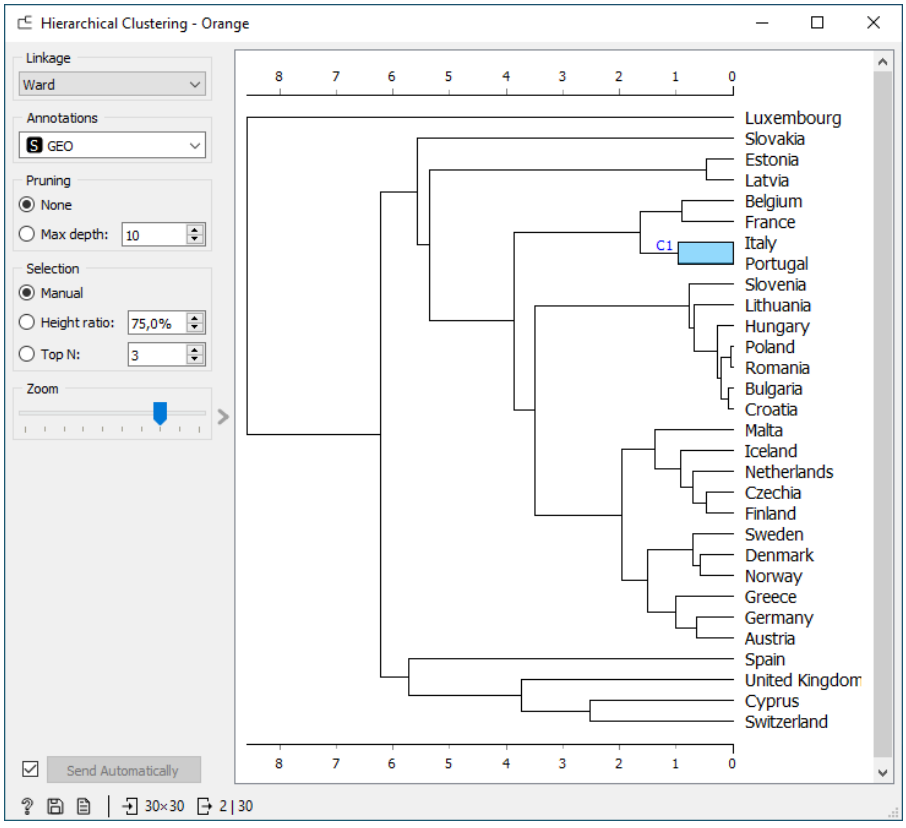
GEO	EU-countries except reporting country	European non-EU countries	Africa	Northern America	obean, Central and South Ame	Asia
Belgium	7.61679	0.592626	1.84595	0.146633	0.190053	1.31425
Bulgaria	0.115259	0.251334	0.00585577	0.00802609	0.00461363	0.114699
Czechia	1.51289	1.71165	0.0431044	0.0840587	0.021934	0.862892
Denmark	2.39707	1.20332	0.43409	0.173207	0.114909	2.19979
Germany	3.38945	1.23303	0.296419	0.116336	0.113033	3.01628
Estonia	0.615988	16.592	0.00952279	0.0337379	0.00843447	0.110464
Greece	2.01045	5.35368	0.260998	0.0731804	0.0258878	1.3961
Spain	4.69445	0.464454	2.35563	0.0965046	4.40878	0.579401
France	2.20963	0.273589	2.53987	0.0743448	0.308823	0.872977
Croatia	0.180116	0.306916	0.00277029	0.0105177	0.00385024	0.01796
Italy	2.00122	1.82683	1.5265	0.025203	0.568053	1.31684
Cyprus	15.923	2.17666	0.536262	0.162272	0.0663772	6.56625
Latvia	0.340525	19.3228	0.00468692	0.0213515	0.00416615	0.0873156
Lithuania	0.0918016	0.543988	0.00311304	0.0111606	0.0018877	0.0551737
Luxembourg	65.6866	4.36348	1.90693	0.572592	0.735357	1.6931
Hungary	0.86345	0.268472	0.0291291	0.0358269	0.0127012	0.244652
Malta	3.07572	0.45223	0.930395	0.0969424	0.0309712	0.463309
Netherlands	2.10459	0.136398	0.574266	0.116241	0.145532	1.12088
Austria	4.99144	4.54282	0.30614	0.112156	0.124431	2.43196
Poland	0.0622467	0.0864772	0.00663993	0.00524871	0.00248997	0.0400899
Portugal	0.929446	0.512396	1.01219	0.0559745	1.14279	0.220937
Romania	0.0500856	0.0453658	0.0114759	0.0038535	0.00197654	0.0572101
Slovenia	0.273553	3.80707	0.00986051	0.0140284	0.0199752	0.0772576
Slovakia	0.395967	0.0744223	0.00665238	0.0137871	0.00357741	0.0373418
Finland	1.17558	0.684671	0.341577	0.0595892	0.0523686	0.794708
Sweden	3.12699	0.848438	0.795718	0.126771	0.223964	2.07012

Obr. 33 Ukázka vstupních dat

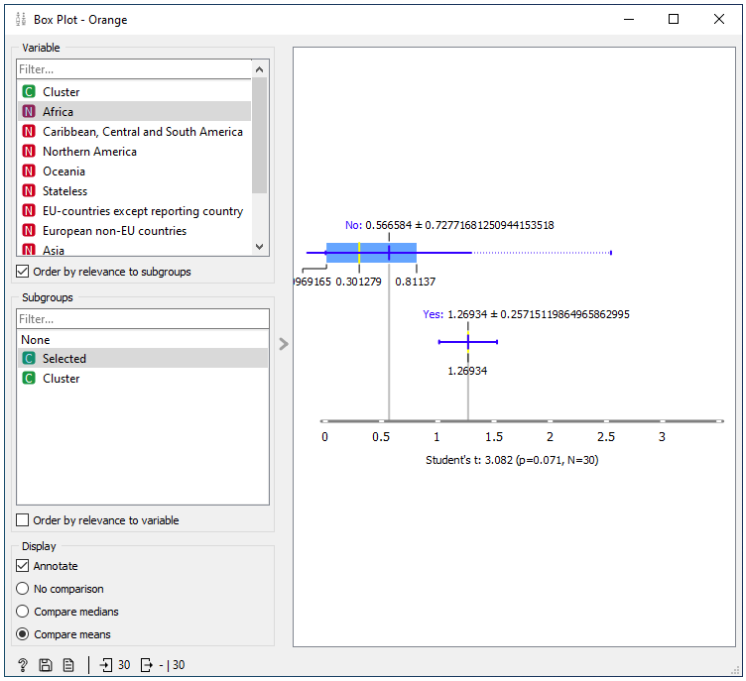
Po připojení uzlu Box Plot se musí změnit tok dat (Reset signals) z hodnoty Selected Data na tok *Data-Data*.

Uzel Box Plot vrací mj. i výsledek dvouvýběrového (nepárového) **Studentova t-testu**, který testuje nulovou hypotézu, že střední hodnota μ_1 jedné skupiny se střední hodnotoa μ_2 jiné skupiny jsou stejná, resp. že rozdíl je konstatní (H_0: $\mu_1 - \mu_2 = konst$).

Proměnné v pravém okně jsou právě seřazeny sestupně (při zaškrtnutí *Oder by relevance to subgroups)* tak, že první vstupní proměnná má nejvyšší hodnotu Studentova t-testu.



Obr. 34 Dendrogram poměrného počtu cizinců žijících v zemích EU



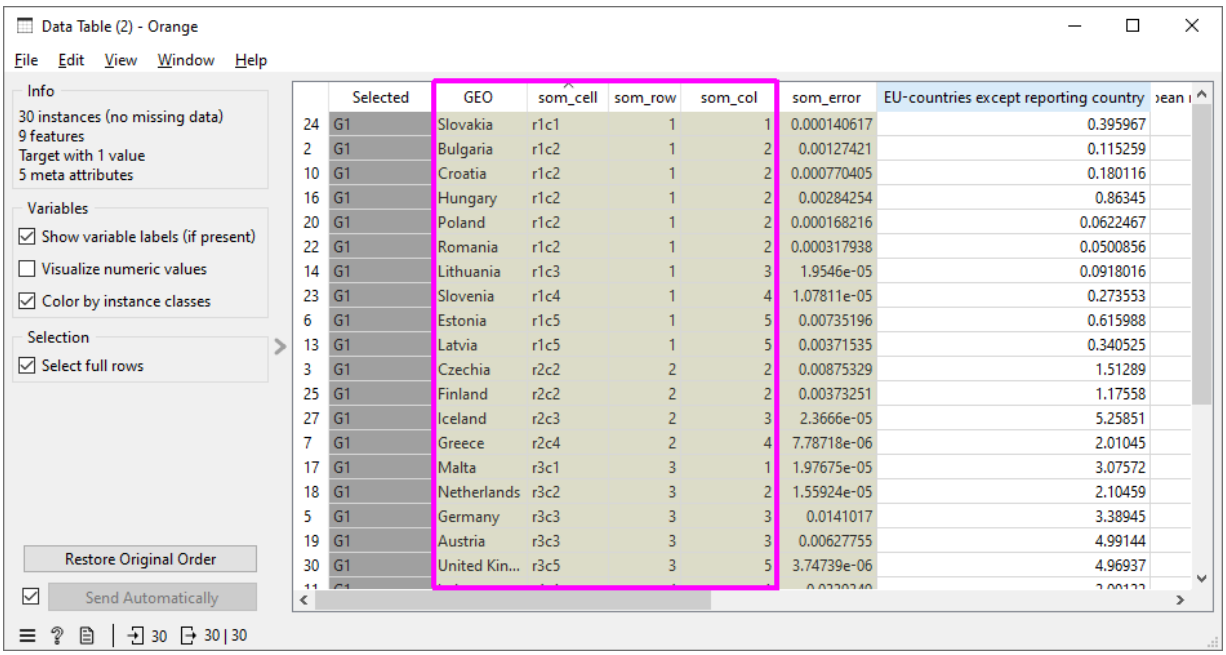
Obr. 35 Boxplot pro shluk C1 (Itálie a Španělsko), kde je vidět, že tento shluk (Yes) má vysoký počet cizinců z Afriky na rozdíl od ostatních států (No)

V dendrogramu vidíme, že Itálie je podobná Portugalsku. Díky boxplotu je viditelné, že tento dvou prvkový shluk se výrazně odlišuje počtem cizinců z Afriky od ostatních států. Atribut Afrika je uveden vlevo jako první a hodnota t-testu t=3 nejvyšší.

Shklukování pomocí Kohononovy mapy – self orgranizing map

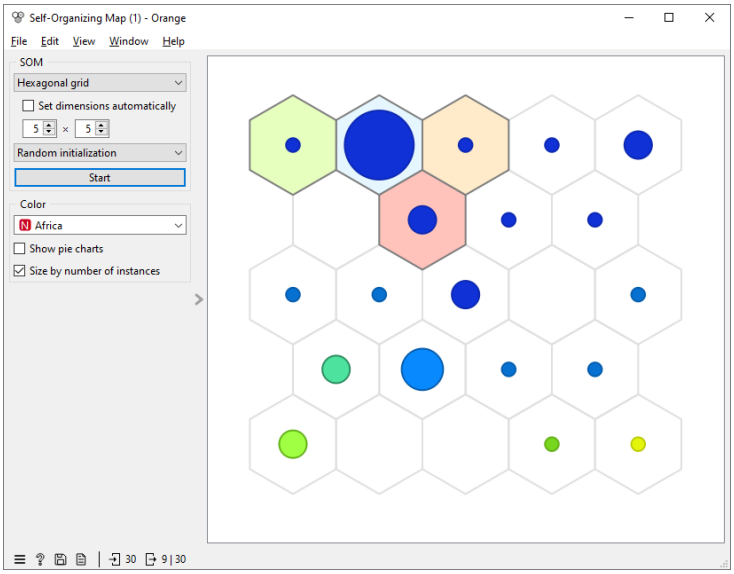
Je zajímavé porovnat výsledky hierarchického shlukování a metody SOM.

Za **uzel SOM** připojte uzel *Data Table*. Při výběru všech buněk myší v uzlu SOM uvidíte výsledné přirazení podobných států do buňky. Lze vybírat i konkrétní buňky v hexagonové mřížce a k nim i přidat okolní buňky a lze tak vidět blízké státy svým procentuálním složením cizinců.



	Selected	GEO	som_cell	som_row	som_col	som_error	EU-countries except reporting country
24	G1	Slovakia	r1c1	1	1	0.000140617	0.395967
2	G1	Bulgaria	r1c2	1	2	0.00127421	0.115259
10	G1	Croatia	r1c2	1	2	0.000770405	0.180116
16	G1	Hungary	r1c2	1	2	0.00284254	0.86345
20	G1	Poland	r1c2	1	2	0.000168216	0.0622467
22	G1	Romania	r1c2	1	2	0.000317938	0.0500856
14	G1	Lithuania	r1c3	1	3	1.9546e-05	0.0918016
23	G1	Slovenia	r1c4	1	4	1.07811e-05	0.273553
6	G1	Estonia	r1c5	1	5	0.00735196	0.615988
13	G1	Latvia	r1c5	1	5	0.00371535	0.340525
3	G1	Czechia	r2c2	2	2	0.00875329	1.51289
25	G1	Finland	r2c2	2	2	0.00373251	1.17558
27	G1	Iceland	r2c3	2	3	2.3666e-05	5.25851
7	G1	Greece	r2c4	2	4	7.78718e-06	2.01045
17	G1	Malta	r3c1	3	1	1.97675e-05	3.07572
18	G1	Netherlands	r3c2	3	2	1.55924e-05	2.10459
5	G1	Germany	r3c3	3	3	0.0141017	3.38945
19	G1	Austria	r3c3	3	3	0.00627755	4.99144
30	G1	United Kin...	r3c5	3	5	3.74739e-06	4.96937

Obr. 36 Výsledek shlukování pomocí SOM s vyznačením souřadnice buňky



Obr. 37 Výběr sousedních buněk v hexagonové mřížce

Úkoly

1. Pomocí boxplotu zdůvodněte, kterou hodnotou proměnné si je podobné Estonsko a Litva.

   ……

2. Který stát je výrazně odlišný od ostatních států (připojuje se jako poslední v dendrogramu)?

   ……

3. Pro shlukování vyzkoušejte i metodu SOM. Porovnejte s výsledky hierarchického shlukování.

4. Vyzkoušejte i analýzu hlavních komponent PCA. Kolik hlavních komponent je nejlépe vybrat?

9 TURISTI A JEJICH ZPŮSOBY UBYTOVÁNÍ VE STÁTECH EU

Databáze Eurostat poskytuje statistiku počtu turistů a počty různých ubytovacích kapacit v jednotlivých zemích EU.

Vstupní data: *Turists_EU.xlsx - list 2019 (resp. list 2023)*

Eurostat kódové označení: **TOUR_OCC_NINAT**, ke stažení z *https://doi.org/10.2908/TOUR_OCC_NINAT*

Cvičná data obsahují údaje o turistech (domácích i zahraničních) v zemích EU z databáze Eurostat za rok 2019 a rok 2023. Podle potřeby lze stáhnout jiný rok nebo aktuální data.

Celkem je v datech 5 údajů pro rok 2019 (vysvětlení viz list Comments):

1. *Turists - Domestic country – počet **domácích** turistů, kteří přenocovali na území státu*
2. *Turists - Foreign country - počet **cizích** turistů, kteří přenocovali na území státu*
3. *Turists Total – součet všech turistů, kteří přenocovali (součet dvou předešlých údajů)*
4. *Establishments for accommodation – počet zařízení určených k přenocování (hotely apod.)*
5. *Bedplaces – celkový počet postelí v konkrétním státě*

Jako opakování sestavte jedno nebo několik workflow, kde budete řešit následující úkoly.

Úkoly:

1. **Chybějící data**
   a. Zjistěte, které hodnoty atributů chybí a v kolika záznamech
   b. Udělejte rozvahu nad imputací a tuto imputaci poveďte, zdůvodněte ji. Dále pokračujte s takto upravenými daty.
   c. Soubor s imputovanými daty odevzdáte jako jeden z výstupů - ***Turists_imp.xlsx***
2. **Korelace** – zjistěte korelaci atributů.
   a. Které korelace jsou nejvyšší a mezi kterými atributy?
   b. Zvažte, zda budete pokračovat se všemi 5 údaji v dalším zpracování.
3. **Určete odlehlé hodnoty –** identifikujte pomocí Boxplotu, Scatter plotu a uzlu Outliers odlehlé hodnoty – data, kde budou identifikované odlehlé hodnoty, uložte jako ***Turists_out.xlsx***
   Sestavte **Line plot** a okomentujte výsledky na základě všech kroků prvotní analýzy.
4. **Proveďte normalizaci dat na interval [0,1],** protože státy EU jsou různě velké rozlohou a počtem obyvatel, dále pracujte s těmito daty, uložte jako ***Turists_norm.xlsx***
5. **PCA** – proveďte analýzu hlavních komponent nad atributy.
   a. Kolik volíte výsledných komponent?
   b. Jaký je vysvětlený rozptyl (variance)?
   c. Soubor s napočítanými hlavními komponentami odevzdáte jako jeden z výstupů ***Turists_PCA.xls***
6. **Shlukování** – realizujte dvě shlukovací metody **K-Means** a **Hierarchické shlukování** (zvolte vhodnou metriku)
   a. Určete výsledný vhodný počet shluků K (zdokumentujte graficky) a shluky uložte do výstupního souboru ***Turists_K.xlsx*** a ***Turists_Hier.xlsx***.
   b. Interpretujte výsledné shluky – které státy jsou si podobné a čím (popis shluku, kvalita shluku – Silhouette plot,…).
   c. Porovnejte výsledky obou shlukovacích metod – v čem se liší a v čem se shodují výsledky, které poskytly.

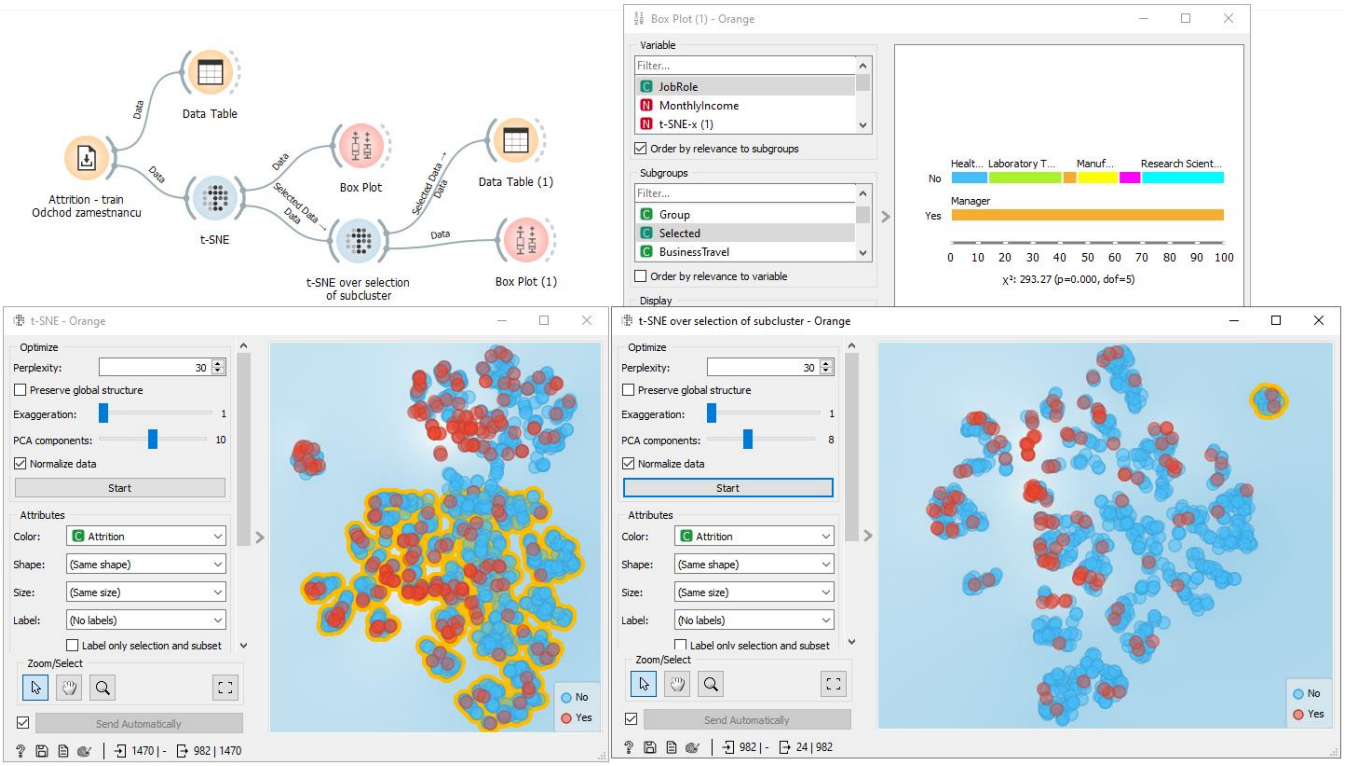
> *Tento příklad je opakováním již získaných znalostí a dovedností.*

10 ODCHOD ZAMĚSTNANCŮ, REDUKCE DIMENZÍ POMOCÍ T-SNE A PREDIKCE

Dataset *Attrition – Train* jsou fiktivní data z Watson Analytic Sample Data od IBM data scientists. Odchody zaměstnanců vedou k zajímavým otázkám typu: „Jaká je limitní vzdálenost dojížďky do zaměstnání podle pozice ve firmě?“ nebo „Porovnej měsíční příjem podle vzdělání a podle toho, zda zaměstnanec odešel?“. Na datasetu bude ukázáno vyhledávání shluků po redukci dimenzí pomocí uzlu **t- SNE** a následném vyšetření pomocí boxplotu.

Dataset Attrition- Train (z data sw Orange)
Workflow *Attrition_TSNE.ows*

Inspirace: https://www.youtube.com/watch?v=fMH0FHj_sww&t=51s



Obr. 38 Vyšetřování dat odchodu zaměstnanců pomocí t-SNE

V uzlu t-SNE lze vybrat shluk a zkoumat jej pomocí boxplotu. První boxplot vyšetřuje malý shluk vlevo nahoře. Je nutné při spojení s boxplotem změnit *Reset Signals* na *Data* místo výchozí volby *Selected Data*.

Zajímavé je zpracování druhým boxplotem. Druhý boxplot vyšetřuje větší shluk uprostřed, který je vybrán a následně ještě jednou analyzován pomocí druhého uzlu t-SNE. Po druhé aplikaci t-SNE dojde k lepšímu rozdělení vybraných dat do shluků. Malý shluk vpravo nahoře ukazuje v boxplotu skupinu osob, které mají JobRole-Manager. Protože jsou atributy seřazeny podle relevance (zatržítku *Order by relevance* je patrně, že druhým určujícím atributem je měsíční příjem *MonthlyIncome*.

Úkoly:

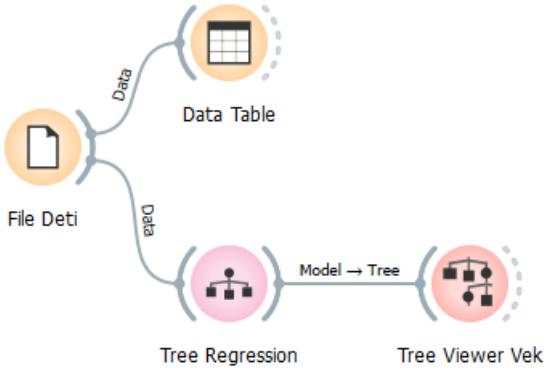
1. Zjistěte, kolik zaměstnanců odešlo z firmy.
2. Které oddělení má nejvíce odešlých zaměstnanců (v procentech).
3. Odešli převážně zaměstnanci, kteří dojíždějí z větší dálky (*DistanceFromHome*)?
4. V datasetech je i soubor *Attrition – Predict* se třemi osobami vytvořený laboratoří Biolab pro výukové potřeby. Zkuste si vyzkoušet různé predikce (*Decission Tree, Random Forest, Naive Bayes, Logistic Regression)*

11 REGRESNÍ ROZHODOVACÍ STROM

Dataset: *3_Deti.xlsx* *RegresníStromDeti.ows (Cvicebnice)* Features: *Vaha, Vyska*; Target: **Vek**

Úkol: Z váhy a výšky predikovat věk dítěte.

Dotaz: Co jsou ta čísla uzlech?



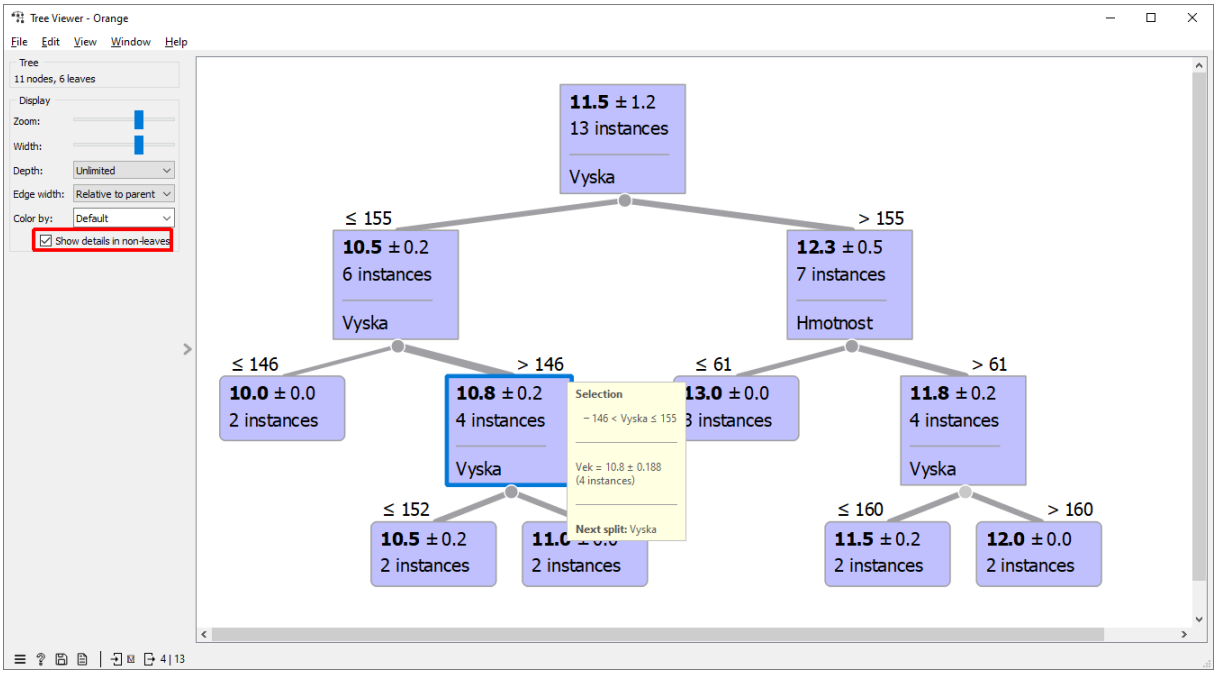
Obr. 39 Workflow pro regresní strom

První řádka je vždy věk a jeho rozpětí, takže v celém datasetu máme děti v průměrném věku 11,5 ± 1,2 roků.

V každém uzlu ukazuje průměrný věk s rozpětím instancí, které reprezentuje uzel.

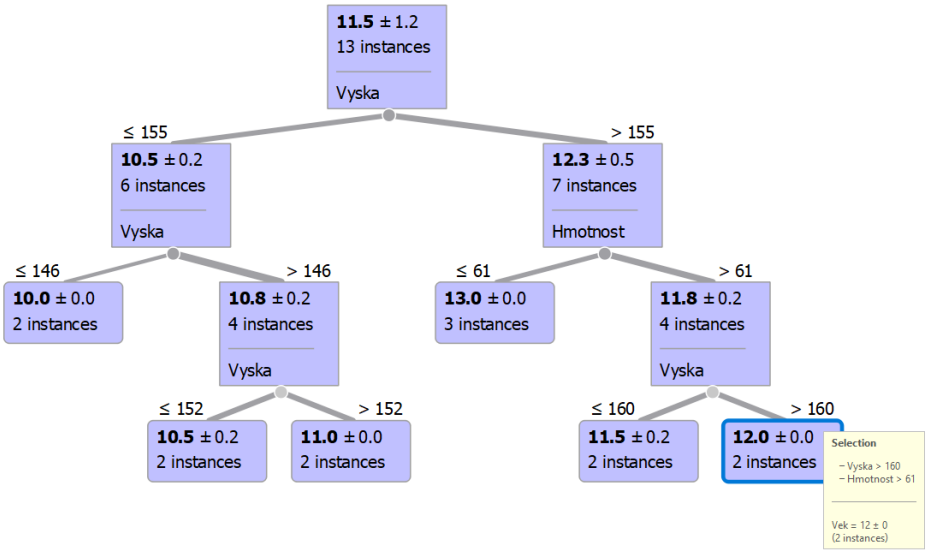
Jinak pravidla se interpretují jako: Když je výška v intervalu <i,j > a hmotnost<>, pak věk je v intervalu ...

Pozn.: Po zatrhnutí vlevo volby Show details… se ukazuje věk v každém uzlu, jinak jen v listech.



Obr. 40 Výsledný regresní strom pro věk dítěte

Např. na následujícím obrázku: *Když je výška > 160 cm a hmotnost > 61 kg, pak věk je 12 roků.*



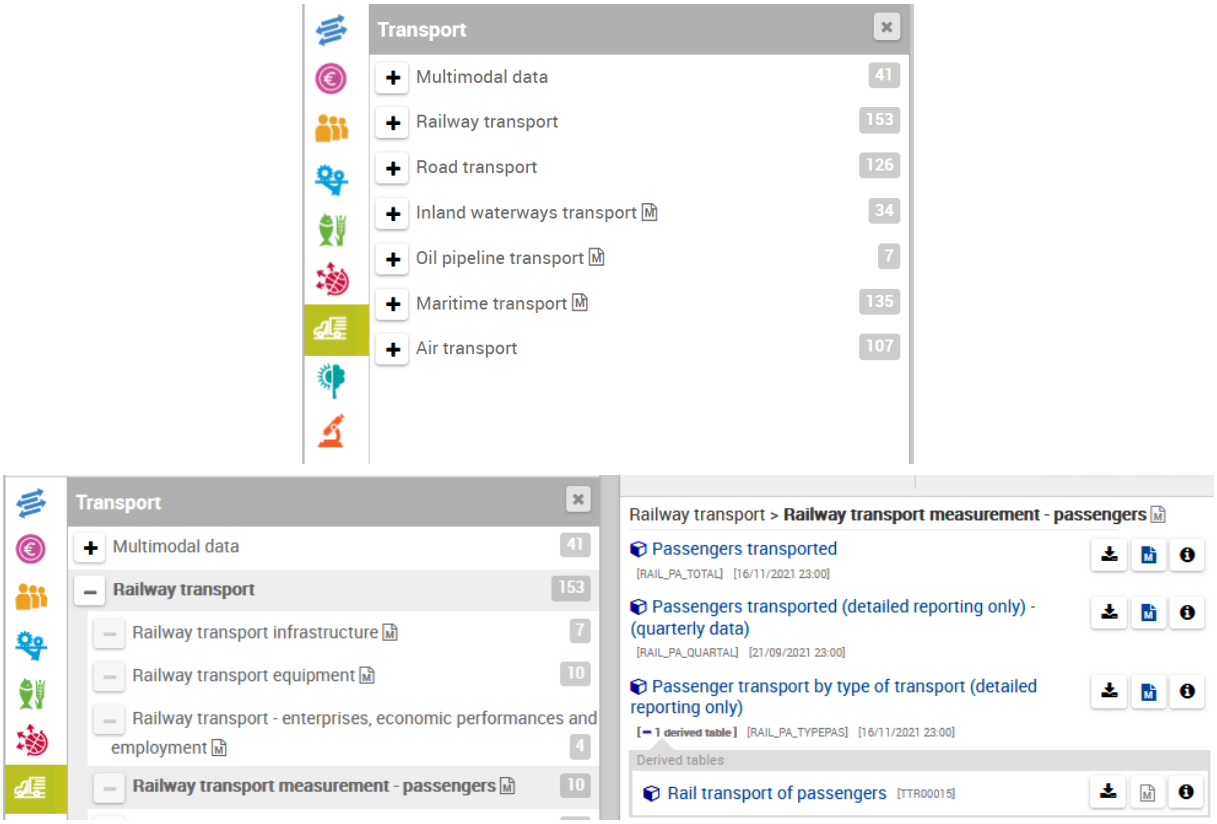
Obr. 41 Pop-up okno, které zobrazuje pravidla a výsledný věk pro dvě instance, které reprezentuje vybraný list

Úkoly

1. Zkuste zaměnit features – výška a věk, a jako target volit např. hmotnost. Interpretujte listy stromu a jednotlivá pravidla.
2. Najděte jiný dataset (třeba EU dat), kde jsou na vstupu číselné hodnoty nebo dichotomické hodnoty a predikujte číselnou hodnotu.
3. Zkuste použít model Random forest a uzel Prediction a porovnejte kvalitu modelů. Můžete zkusit predikovat věk pro nové dítě o výšce 150 cm a váze 53 kg.

12 ČASOVÉ ŘADY ŽELEZNIČNÍ DOPRAVY STÁTŮ EU

Databáze Eurostat poskytuje v sekci *Transport* údaje o infrastruktuře, přepravních výkonech a nehodách v dopravě. Údaje jsou rozčleněny podle druhů dopravy. V sekci železniční doprava lze nalézt údaje o vybavenosti tratí, jako jsou počty lokomotiv, délka tratí podle trakce a rychlosti. Přepravní výkony jsou sledovány v objemu přepraveného zboží a přepravených osob.



Obr. 42 Sekce Transport databáze Eurostat

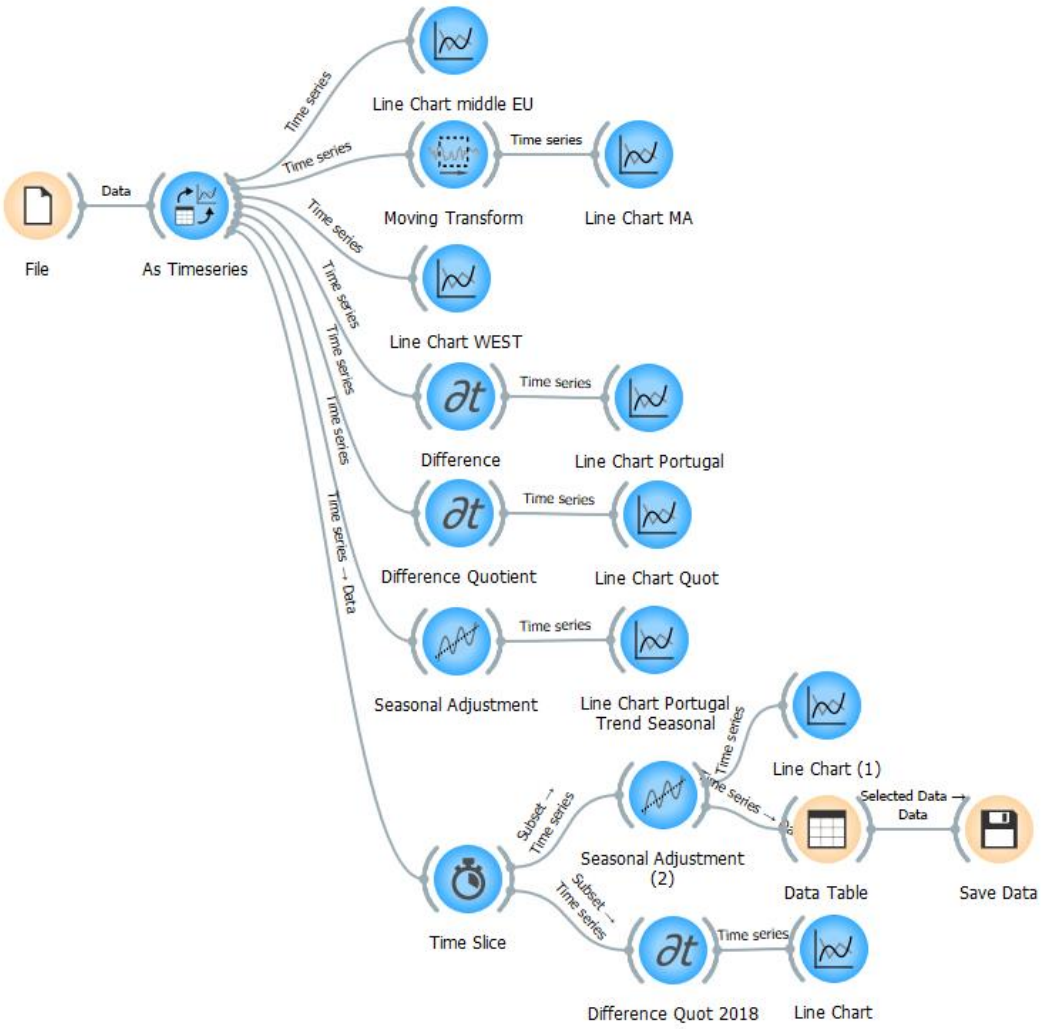
Vzhledem k pandemickým opatřením v roce 2020 a 2021 je zajímavé analyzovat časovou řadu **přepravených osob** na železnici jednotlivých států. Data jsou dostupná od roku 2004 v podrobnosti za jednotlivá čtvrtletí každého roku Q1, Q2, Q3 a Q4. Některé řady nejsou úplné, chybí začátek nebo konec, někdy jen několik hodnot. Měřenou jednotkou jsou tisíce přepravených osob a pro osobokilometry jsou to miliony osobokm. Pro zpracování je přidán sloupec s časovým údajem dne čtvrtletí, tak aby data bylo možné zpracovat v Orange. Pro zpracování v MS Excel je a doplněn složený kód v syntaxi <*Rok Qx*>. Data obsahují i údaje za roky 2020 a část roku 2021, kdy byla v jednotlivých státech omezení pohybu osob dána vládními nařízeními.

Kódové označení statistiky v databázi Eurostat je RAIL_PA_QUARTAL.

Data *rail_pa_quartal_custom.xlsx, rail_pa_quartal_paskm.xlsx (osobokilometry)*

Ke stažení
https://ec.europa.eu/eurostat/databrowser/view/RAIL_PA_QUARTAL/default/table?lang=en&category=rail.rail_pa

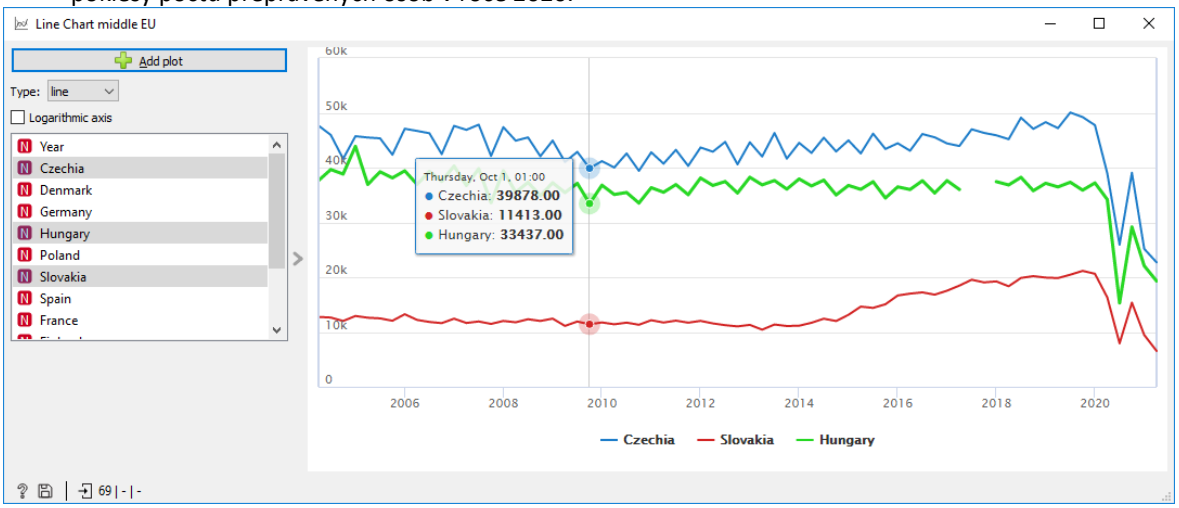
> *Příklad vychází ze článku* (Dobešová et al. 2022).



Obr. 43 Workflow pro zpracování časových řad

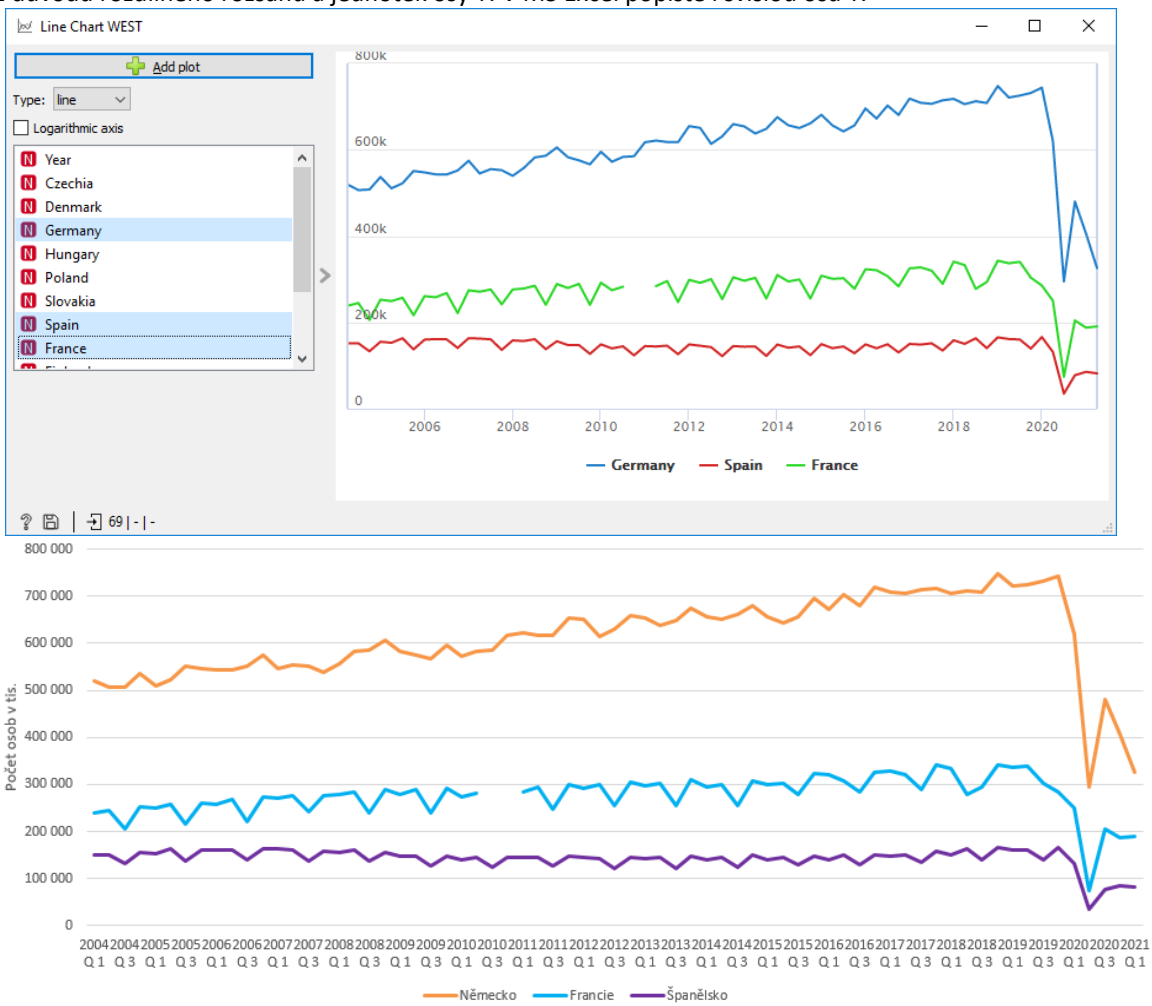
Úkoly

1. Vykreslete časovou řadu pomocí uzlu *Line Chart* pro *Českou republiku, Slovensko, Maďarsko*. Sledujte poklesy počtu přepravených osob v roce 2020.



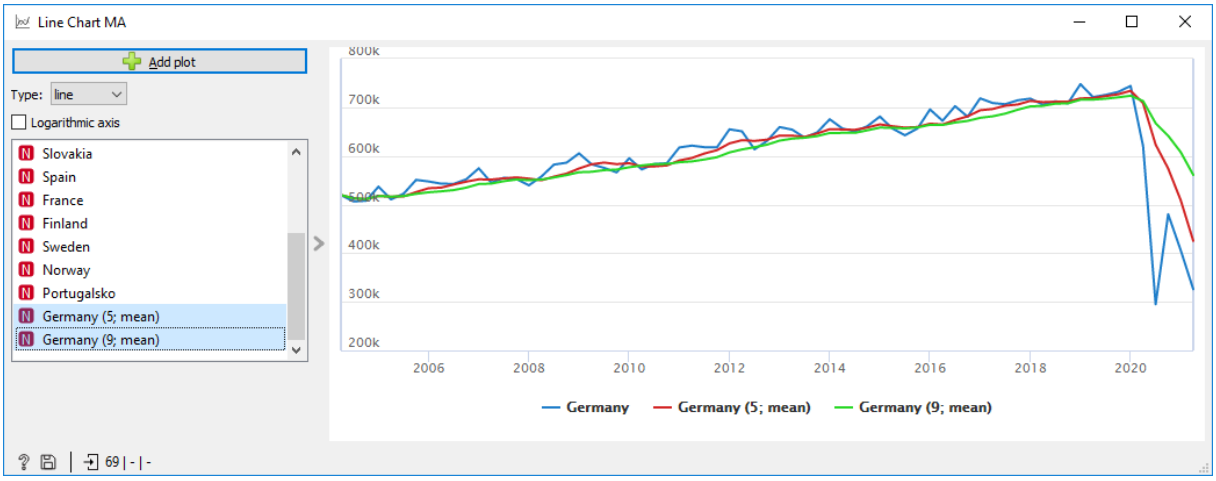
Obr. 44 Časová řada počtu přepravených osob na železnici v České republice, na Slovensku a v Maďarsku (2004-2021)

2. Vykreslete časovou řadu pomocí uzlu v samostatném uzlu *Line Chart* pro Německo, Francii a Španělsko z důvodu rozdílného rozsahu a jednotek osy Y. V MS Excel popište i svislou osu Y.



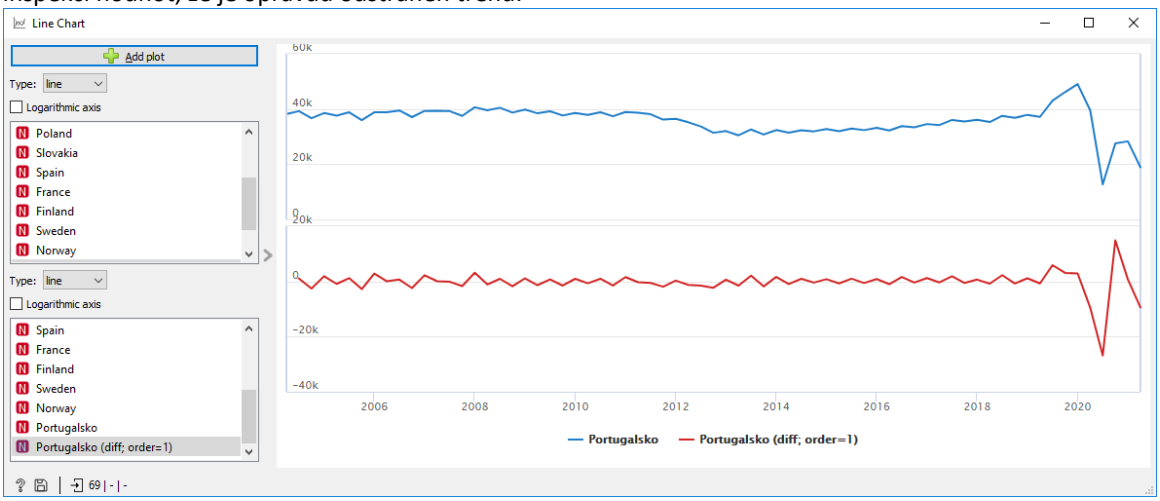
Obr. 45 Časová řada počtu přepravených osob na železnici v Německu, Francii a Španělsku (2004-2021) v Orange (nahoře), v MS Excel (dole)

3. Pomocí uzlu *Moving Transform* vykreslete **trend** časové řady pro Portugalsko, Německo a další státy. Experimentujte s délkou okna pro MA (5 nebo více). Kdy trend stoupá a kdy klesá? Sledujte změnu trendu okolo a po roce 2008, kdy byla hospodářská krize.
4. Rozhodněte, které státy nebudete zpracovávat z důvodu chybějících úajů, krátké časové řady (chybí údaje na začátku řady nebo na konci). Které jsou státy?



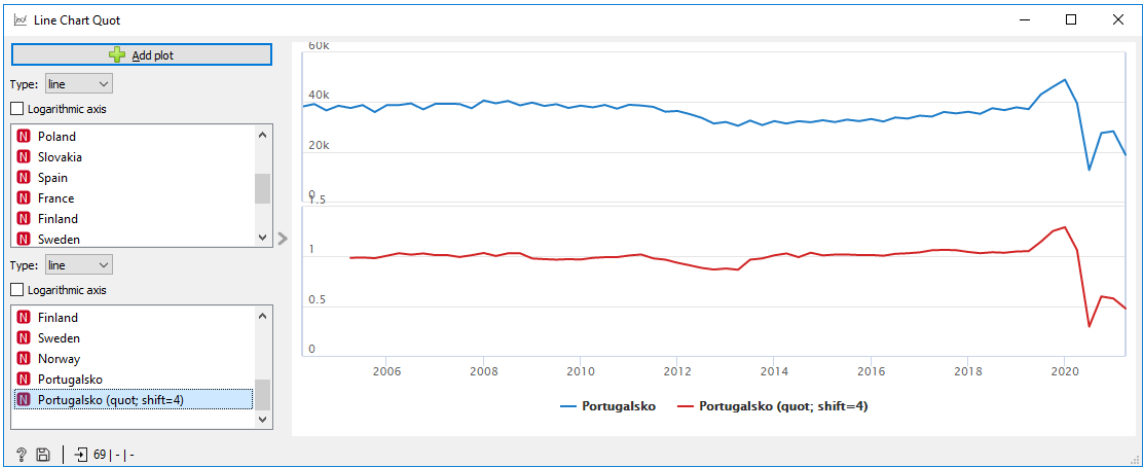
Obr. 46 Časová řada počtu přepravených osob a její trend s různou délkou okna

5. Proveďte výpočet **prvních diferencí** pomocí uzlu *Difference* pro Portugalsko a vykreslete v grafu. Ověřte inspekcí hodnot, že je opravdu odstraněn trend.



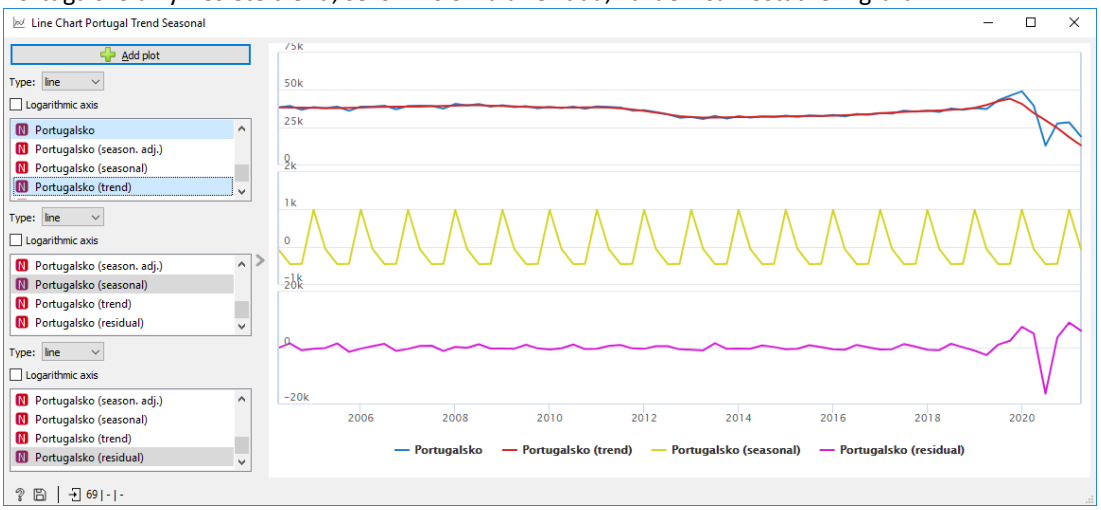
Obr. 47 Časová řada a její první diference

6. Spočítejte **temto růstu** pomocí uzlu *Difference (*volba *quotient, shift 4)* pro Portugalsko. Jaké jsou relativní změny dvou čtvrtletí v následujících rocích? Kdy byl maximální pokles? Kdy byla doba růstu následovaná dobou poklesu?



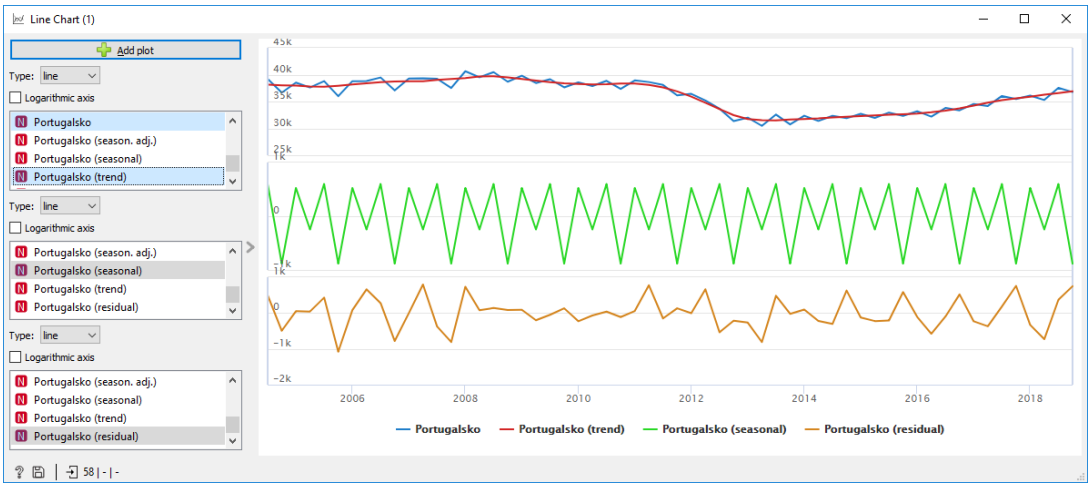
Obr. 48 Časová řada a tempo růstu pro Portugalsko

7. Proveďte **aditivní rozklad** časové řady pomocí uzlu *Seasonal Adjustment* s délkou periody 4 pro Portugalsko a vykreslete trend, sezonní složku a rezidua, každé v samostatném grafu.



Obr. 49 Aditivní rozklad časové řady počtu cestujících v Portugalsku od roku 2004 do roku 2021

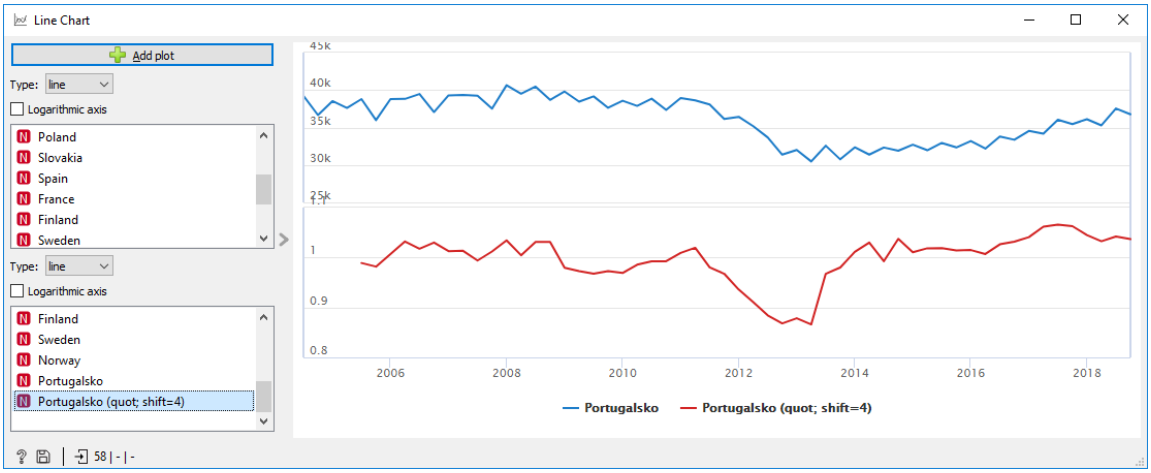
8. Pomocí uzlu *Time Slice* zkraťte časovou řadu od roku 2004 jen do roku 2018 a následně proveďte stejný **rozklad** pomocí *Seasonal Ajustment* jako v předchozím úkolu. Porovnejte, jak se zejména změnila sezónní složka a rezidua.



Obr. 50 Aditivní rozklad zkrácené časové řady počtu cestujících v Portugalsku od roku 2004 do roku 2019

9. Spočítejte **temto růstu** pomocí uzlu *Difference (*volba *quotient, shift 4)* pro zkrácenou osu, kde budou lépe viditelné relativní změny v intervalu <-1, 1>. Tempo růstu vždy zobrazte v samostatném grafu, kdy křivka osciluje kolem hodnoty 1.

Tempo růstu mezi čtvrtletími se spočítá podle vzorce $k = y_t / y_{t-4}$.



Obr. 51 Tempo růstu zkrácené řady pro Portugalsko

10. Vyhodnoťte, která evropský stát měl nejvyšší pokles tempa růstu v roce 2020 oproti roku 2019.
11. Vypočítejte tempa růstu s hodnotou *shift* 8 a vyhodnoťte pokles tempa růstu roku 2020 oproti roku 2018 a také zjistěte změnu tempa růstu roku 2021 oproti 2019 (tzn. porovnává se epidemický rok 2021 oproti roku 2019 před epidemií). Má některý stát tempo růstu (se shift 8) v roce 2021 vyšší než jedna?
12. Spočítejte *průměrný koeficient tempa růstu* pro rozmezí 2004 až 2019 podle vzorce

$$\bar{k} = \sqrt[n-1]{k_2 * k_3 * \ldots * k_n}$$

a porovnejte, jak se liší průměrné tempo od tempem růstu v roce 2020.
Pro výpočet použijte MS Excel.
13. Všechny získaná data (diference, tempa růstu, trendy, sezónní složky, rezidua) uložte so souborů MS Excel.

13 LITERATURA

DOBEŠOVÁ, Zdena, 2022. *ORANGE, Praktický návod do cvičení předmětu Data Mining* [online]. Olomouc, Czech Republic: Univerzita Palackého v Olomouci. ISBN 978-80-244-6086-4. Dostupné z: doi:10.5507/prf.22.2440864

DOBEŠOVÁ, Zdena, Karel MACKŮ a Michal KUČERA, 2022. Výuka geoinformatických předmětů na příkladech dat Evropské unie. In: *Sympozium GIS Ostrava*.

EUROPEAN COMMISSION, 2008. *NACE Rev. 2 Statistical classifcation of economic activites in the European Community* [online]. Methodolog. Luxembourg: Luxembourg: Office for Official Publications of the European Communities. ISBN 978-92-79-04741-1. Dostupné z: https://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/ks-ra-07-015

EUROSTAT, 2022. *NACE Rev. 2 - Statistical classification of economic activities* [online] [vid. 2021-02-10]. Dostupné z: https://ec.europa.eu/eurostat/web/nace-rev2

MASOPUST, Jan, Zdena DOBESOVA a Karel MACKŮ, 2021. Utilisation of EU Employment Data in Lecturing Data Mining Course BT - Artificial Intelligence in Intelligent Systems. In: Radek SILHAVY, ed. Cham: Springer International Publishing, s. 601–616. ISBN 978-3-030-77445-5.

PÁSZTO, Vít, Andreas REDECKER, Karel MACKŮ, Carsten JÜRGENS a Nicolai MOOS, 2020. Data Sources. In: Vít PÁSZTO, Carsten JÜRGENS, Polona TOMINC a Jaroslav BURIAN, ed. *Spationomy: Spatial Exploration of Economic Data and Methods of Interdisciplinary Analytics* [online]. Cham: Springer International Publishing, s. 3–38. ISBN 978-3-030-26626-4. Dostupné z: doi:10.1007/978-3-030-26626-4_1

PRETNAR, Ajda, 2021. *Box Plot Alternative: Violin Plot* [online] [vid. 2021-08-05]. Dostupné z: https://orangedatamining.com/blog/2021/2021-08-05-violin-plot/