Ekonomická fakulta
Faculty of Economics

Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Jihočeská univerzita v Českých Budějovicích
Ekonomická fakulta
Katedra aplikované matematiky a informatiky

Bakalářská práce

Zdroje ekonomických dat ve světě a možnosti jejich zpracování programem R

Vypracoval: Marek Svoboda
Vedoucí práce: Ing. Michael Rost, Phd.

České Budějovice 2018

JIHOČESKÁ UNIVERZITA V ČESKÝCH BUDĚJOVICÍCH
Ekonomická fakulta
Akademický rok: **2017/2018**

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

(PROJEKTU, UMĚLECKÉHO DÍLA, UMĚLECKÉHO VÝKONU)

Jméno a příjmení:	**Marek SVOBODA**
Osobní číslo:	**E14380**
Studijní program:	**B6209 Systémové inženýrství a informatika**
Studijní obor:	**Ekonomická informatika**
Název tématu:	**Zdroje ekonomických dat ve světě a možnosti jejich zpracování programem R**
Zadávající katedra:	**Katedra aplikované matematiky a informatiky**

Zásady pro vypracování:

Práce se zaměří na vytvoření studie popisující světové zdroje ekonomických dat, jejich získání a způsoby další manipulace s nimi. Vedle veřejných zdrojů budou zkoumány i zdroje s omezeným přístupem.
V aplikační části se zpracovatel věnuje otázce manipulace se získanými daty pro potřeby využití software R (různé formáty dat, chybějící data, transformace dat atd.).

Metodický postup:

1. Studium relevantní literatury.
2. Investigativní činnost na Internetu:
   a) obecné zdroje: World Bank Data, United Nations Data, WTO Data, OECD Statistics, atd.
   b) národní zdroje: Český statistický úřad, UK government Data Project, US Government Data Project, Eurostat's Databases, US Bureau of Economic Analysis.
   c) další zdroje: Thomson Reuters Datastream, Infochimps.
3. Zkoumání otázek spojených se zpracováním dat pro program R.
4. Sestavení a uspořádání vhodných ukázek.
5. Závěr a doporučení.

Rozsah grafických prací: **dle potřeby**

Rozsah pracovní zprávy: **40 - 50 stran**

Forma zpracování bakalářské práce: **tištěná**

Seznam odborné literatury:

1. **Chambers, J. M. (2008).** *Software for Data Analysis: Programming with R.* **New York: Springer.**
2. **Spector, P. (2008).** *Data Manipulation with R.* **New York: Springer.**
3. **Warnes, G. R.** *gdata: R Programming Tools for Data Manipulation.* **R pagkage version 1.4-6.**
4. **Williams, G. (2011).** *Data Mining with Rattle and R.* **Springer, New York.**

Vedoucí bakalářské práce: **Ing. Michael Rost, Ph.D.**
Katedra aplikované matematiky a informatiky

Datum zadání bakalářské práce: **27. února 2018**

Termín odevzdání bakalářské práce: **15. dubna 2018**

doc. Ing. Ladislav Rolínek, Ph.D.
děkan

RNDr. Jana Klicnarová, Ph.D.
vedoucí katedry

V Českých Budějovicích dne 27. února 2018

Prohlášení

Prohlašuji, že svoji bakalářskou práci jsem vypracoval samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47 zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své bakalářské práce, a to - v nezkrácené podobě/v úpravě vzniklé vypuštěním vyznačených částí archivovaných Ekonomickou fakultou - elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

V Českých Budějovicích, dne 11. 6. 2018 Podpis studenta……..

Poděkování

Tímto bych rád poděkoval vedoucímu mé bakalářské práce, panu Ing. Michaelovi Rostovi, Phd. za pomoc, ochotu a věnovaný čas při zpracování této bakalářské práce.

Obsah

1 Úvod

Informace jsou v moderní době velice důležitou součástí života každého člověka. Díky nim můžeme činit správné a podložené rozhodnutí, což nám přináší užitek. Jednou z možností, jak získat informace, je nalezení dat, jejich zpracování a vizualizace a následné pochopení výstupů zpracovaných dat. Zmíněná data je ovšem nutné získat z nějakého zdroje. Tyto zdroje mohou mít formu fyzickou, ať už se jedná o knihy, časopisy, studie či jiné listiny, anebo formu digitální v podobě databází a internetových stránek. Těchto digitálních zdrojů, které jsou dostupné na internetu, je však velké množství a každý se zaměřuje na jinou skupinu uživatelů a dat, a proto není snadné najít ten správný zdroj, který vyhovuje konkrétnímu uživateli.

Cílem této práce je proto nalezení a popsání určité skupiny těchto zdrojů. Touto skupinou jsou zdroje, které poskytují ekonomická data. Tuto skupinu jsem vybral z důvodu svého studia na obchodní akademii a poté na ekonomické fakultě a osobní zkušeností s tím, že není snadné vyhledat validní a relevantní ekonomická data pro různé prezentace a studentské práce. Touto bakalářskou prací bych chtěl popsat možnosti získávání dat z různých zdrojů a tím ulehčit hledání relevantního zdroje pro budoucí studenty. Dalším cílem je zpracování ekonomických dat. K tomuto účelu existuje mnoho softwarů, v této práci je použito programovací prostředí R. Tento program jsem vybral z důvodu jeho všestrannosti, dostupnosti a také z důvodu rozšíření znalostí.

Práce je rozdělena do dvou částí. V první, teoretické části, je popsáno programovací prostředí R, programovací jazyk R a jeho syntaxe a vybrané zdroje ekonomických dat a možnosti získávání dat pomocí těchto zdrojů. Druhá část je praktická. Zde je ukázána práce se čtyřmi knihovnami programu R (Quandl, wbstats, eurostat, WDI), jejichž primární využití je získání dat z různých zdrojů pomocí příkazů programovacího jazyka R. Dále je v práci ukázka zpracování dat získaných pomocí dříve zmíněných knihoven v programu R.

2 Teoretická část

Tato kapitola obsahuje popis zdrojů ekonomických dat a možnosti získávání dat pomocí různých nástrojů. Dále je zde představen program R, jeho prostředí a syntaxe programovacího jazyka, který je zde využíván.

2.1 Zdroje ekonomických dat

V této podkapitole je popsána investigativní činnost, která spočívá v hledání zdrojů ekonomických dat na internetu a jejich popisem. U každého zdroje je nejdříve stručně popsán provozovatel daného zdroje. Poté následuje shrnutí možností získávání dat u každého zdroje.

2.1.1 World Bank Data

Světová banka (World Bank) je termín označující dvě organizace, které patří pod Organizaci spojených národů. Těmito organizacemi jsou Mezinárodní banka pro obnovu a rozvoj (International Bank for Reconstruction and Development – IBRD) a Mezinárodní asociace pro rozvoj (International Development Association – IDA). V současné době je členem Skupiny Světové banky 187 zemí. Získávání dat z tohoto zdroje je zdarma. (Světová banka, 2018)

Získávání dat Světové banky

Databáze dat Světové banky se nachází na internetové adrese http://data.worldbank.org/. Tento zdroj umožňuje uživateli získávat data ze 187 zemí světa, které jsou navíc rozdělené do skupin dle různých atributů jako je poloha, úroveň příjmů, velikost území, členství v organizacích a mnoho dalších. O těchto zemích je zde uveřejněno přes sedmnáct tisíc ukazatelů v odvětví zemědělství, ekonomiky, vzdělání, vědy, demografie a dalších, některá data jsou k dispozici již od roku 1960. Tyto informace jsou prezentovány výborně zpracovaným grafickým rozhraním a uživatel má možnost interakce s daty. Světová banka poskytuje všechna data pro nekomerční využití, a to ve formátech CSV, XML a XLS.

Pro vývojáře je poskytována služba API. Každý dotaz na začátku musí obsahovat **http://api.worldbank.org/v2/“** nebo **„https://api.worldbank.org/v2/**. Dále má uživatel na výběr ze dvou možností struktury dotazu, a to argumentační struktura nebo URL struktura. Například pro vytvoření dotazu pro získání zemí s nízkou úrovní příjmů může uživatel k povinnému začátku dotazu přidat **countries?incomeLevel=LIC**

nebo **incomeLevels/LIC/countries** a v obou případech dostane stejný výsledek. Pro doplnění dalších parametrů dotaz musí pokračovat symbolem „**?**“. Dostupné parametry jsou datum, formát (JSON, XML, JSONP), stránka, počet položek na jedné stránce, MRV nebo MRNEV (zobrazí nejaktuálnější hodnoty parametru na základě hodnoty tohoto parametru), frekvence (měsíční, čtvrtletní, roční). Kompletní dokumentace k API světové banky je k dispozici na webu Světové banky. (Světová banka, 2018)

Práce s knihovnami programu R wbstats a WDI, které zpřístupňují službu API Světové banky pro syntaxi programovacího jazyka R, je ukázána v kapitolách 3.2 a 3.4 této bakalářské práce.

2.1.2 United Nations Data

United Nations Data (česky data Organizace spojených národů) je zdroj dat provozován Statistickou divizí organizace spojených národů patřící pod Oddělení ekonomiky a sociálních záležitostí. Tato služba byla spuštěna v roce 2005 v rámci projektu nazvaným „Statistics as a Public Good“ (Statistika jako veřejný statek), jehož cílem je poskytnout volný přístup k světovým statistikám, poučit uživatele o důležitosti statistiky při rozhodování na základě faktů a asistovat národním statistickým úřadům členských zemí při rozšiřování jejich statistických údajů. Data jsou poskytována zdarma. (Organizace spojených národů, 2018)

Získávání dat OSN

Internetová adresa databáze tohoto zdroje je http://data.un.org/. Vyhledáním požadovaného ukazatele se zobrazí tabulka obsahující jeho hodnoty pro všechny země a roky, které jsou dostupné. Uživatel má možnost tato data filtrovat dle vlastních potřeb a poté uložit pro vlastní potřebu pro jejich nekomerční využití. Dostupné formáty pro stažení jsou XML a CSV. Jsou zde dostupná data od roku 1960 až po současnost pro 220 států.

Služba API poskytovaná Organizací spojených národů je založena na standardu SDMX-RI (SMDX Reference Infrastructure) od Eurostatu a využívá službu SOAP. Uživatel nejdříve musí vytvořit dotaz SDMX, což je dokument XML, který definuje parametry dotazu. Je možné jej jednoduše vytvořit na stránkách http://data.un.org/SdmxBrowser/start. Poté je nutné tento dotaz vložit do webové služby SOAP nacházející se na stránce http://data.un.org/ws/NSIStdV20Service.asmx. Výsledek dotazu je uživateli poskytnut v souboru XML, který pak je možné importovat

do různých statistických programů a zpracovat ho. Více informací o této službě je dostupných na adrese webových stránkách této organizace.

2.1.3 IMF Data

Mezinárodní měnový fond (International Monetary Fund – IMF) je organizace 189 zemí, které podporují světovou měnovou spolupráci, zajišťují finanční stabilitu, usnadňují mezinárodní spolupráci, podněcují vysokou zaměstnanost a udržitelný ekonomický růst a redukují chudobu po celém světě. Vznikla v roce 1944 na konferenci v New Hampshiru. Jejím hlavním cílem je zajistit stabilitu mezinárodního měnového systému, což je systém směnných kurzů a mezinárodních plateb, který umožňuje platby mezi jednotlivými státy. Data získaná z tohoto zdroje jsou bez poplatku. (Mezinárodní měnový fond, 2018)

Získávání dat IMF

Webová adresa tohoto zdroje je http://data.imf.org/. Zde se nachází seznam databází, které jsou sem nahrávány každý rok v dubnu a v říjnu. Pro uživatele jsou zde dostupná data od roku 1980 až po současnost a nachází se zde i odhady pro roky budoucí. Pro vlastní využití je možnost stáhnutí dat ve formátu ASPX, který je kompatibilní s programem Microsoft Excel, R a mnoha jinými statistickými programy.

Služba API, kterou IMF poskytuje, je založená na standardu SDMX. Vytváření dotazů v tomto standardu je popsáno v kapitole 2.1.2 této práce. IMF navíc poskytuje možnost pracovat s formátem JSON. Její použití je detailně popsáno na internetové stránce této organizace.

2.1.4 WTO Data

Světová obchodní organizace (World Trade Organization – WTO) je organizace 164 států založená v roce 1995 po osmiletém vyjednávání zvaném Uruguayské kolo. Jejím cílem je vytvoření a udržení mnohostranného obchodního systému. Mezi principy, kterými se tento systém řídí, patří žádná diskriminace, odstraňování bariér obchodu a nekalých praktik, transparentnost, pomoc méně rozvinutým státům a ochrana životního prostředí. WTO poskytuje data zdarma. (Světová obchodní organizace, 2018)

Získávání dat WTO

Získávání dat z tohoto zdroje je možné na internetové adrese http://stat.wto.org. Zde je pro uživatele k dispozici výběr z několika prezentací dat. Prvním jsou **Trade Profiles** (obchodní profily), které poskytují standardní informace o obchodní situaci a obchodní politice více než sto osmdesáti členů, pozorovatelů a jiných vybraných ekonomik. Tyto profily jsou doplněné o základní makroekonomické ukazatele a jsou dostupné ke stažení ve formátech HTML a EXE (verze pro PDF a Excel). Další možnou prezentací dat jsou **Tarrif Profiles** (tarifní profily). Zde jsou data prezentováno pro každou ekonomiku s rozdělením podle sektorů a rozsahu. Souhrnné tabulky (jsou dostupné pouze pro stažení) umožňují porovnávání různých profilů. Dostupné formáty pro stažení jsou opět HTML a EXE (PDF a Excel verze). Dále je možná prezentace pomocí uživatelem vytvořené časové řady. Parametry, které jsou dostupné pro vytvoření těchto řad, jsou předmět obchodu, sledovaná ekonomika, obchodní partneři, import/export, jednotky a roky. Výslednou časovou řadu je možné vidět přímo na internetových stránkách nebo si jí může uživatel stáhnout ve formátech XLS, CSV a XML. Služba API není Světovou obchodní organizací poskytována.

2.1.5 OECD Data

Organizace pro ekonomickou spolupráci a rozvoj (Organization for Economic Co-operation and development – OECD) je organizace třiceti pěti států založená v roce 1961. Jejím hlavním cílem je podporovat zásady, které zlepší ekonomickou a sociální situaci lidí po celém světě. Data jsou poskytována zdarma. (Organizace pro ekonomickou spolupráci a rozvoj, 2018)

Získávání dat OECD

Internetová adresa tohoto zdroje je http://stats.oecd.org/. Zde si může uživatel vyhledat požadovaná data, která jsou mu primárně prezentována pomocí tabulky. U každé tabulky jsou v pravé části stránky informace o zobrazovaném ukazateli (metoda počítání, jednotky atd.). Každou tabulku lze filtrovat podle několika kritérií. Dále má uživatel možnost zobrazovaná data vykreslit do grafů, které se zobrazí přímo v internetovém prohlížeči (je zapotřebí program Adobe Flash Player). Data je také možné stáhnout pro vlastní potřebu ve formátech XLS, CSV nebo DOC. V záložce pro stažení dat je také možné zobrazit dotaz pro službu API ve standardu SDMX, samotná

data jsou pak pomocí této služby předána ve formátu XML. Přihlášený uživatel má dále možnost ukládat dotazy pro data do oblíbených ve svém profilu.

Výše zmiňovaná služba API je také dostupná ve standardu SDMX-JSON, kde se uživateli předávají data ve formátu JSON. Každý dotaz této služby začíná prefixem **http://stats.oecd.org/SDMX-JSON/data/**. Dále dotaz obsahuje kód data setu, filtrování, název agentury spravující data a další nepovinné parametry (časové vymezení, rozsah informací o datech atd.). Každý z těchto parametrů je oddělen lomítkem. Více informací o službě API od OECD je dostupné na internetové adrese tohoto zdroje. (Organizace pro ekonomickou spolupráci a rozvoj, 2018)

2.1.6 CIA World Factbook

Tento zdroj je spravován Ústřední zpravodajskou službou Spojených států Amerických (Central Intelligence Agency – CIA). Tato organizace vznikla v roce 1947 prezidentem Harry S. Trumanem. Část jejich internetových stránek zvaná „The World Factbook“ poskytuje informace o historii, populaci, vládě, ekonomice, geografii, armádě a dopravě pro 267 států a skupin států. (Ústřední zpravodajská služba USA, 2018)

Získávání dat The World Factbook

Internetová adresa tohoto zdroje je https://www.cia.gov/library/publications/the-world-factbook. „The World Factbook“ se od ostatních zdrojů dat popsaných v této bakalářské práci odlišuje tím, že neposkytuje čistá data, ale spíše shrnutí informací o státech ve výše zmíněných oblastech podobnou formou, jakou jsou data prezentována na oblíbených stránkách wikipedia.com. Zdroj přesto obsahuje hodnoty základních ukazatelů v jednotlivých oblastech a jejich nejaktuálnější hodnoty. Archiv na této stránce obsahuje starší verze těchto stránek až do roku 2000, které jsou volně ke stažení.

2.1.7 Wolfram|Alpha

Wolfram|Alpha je výpočetní znalostní systém nebo také odpovídací systém vyvinutý firmou Wolfram Research, která byla založena Stephenem Wolframem. Byl spuštěn v květnu 2009. Veřejnosti je přístupný jako internetová služba, která zodpovídá faktické dotazy vypočítáním odpovědi z externích zdrojů dat. Je založen na Wolfram Mathematica, což je výpočetní nástroj zahrnující počítačovou algebru, symbolické a

numerické výpočty, vizualizaci a statistické metody. Data čerpá z akademických a komerčních internetových zdrojů. (Wikimedia Foundation, Inc, 2018)

Získávání dat Wolfram|Alpha

Internetová adresa tohoto zdroje je http://www.wolframalpha.com/. Data z tohoto zdroje dat získá uživatel pouhým zadáním dotazu (může být i heslovitou formou) v anglickém jazyce na dříve zmíněné internetové stránce. Wolfram|Aplha pak uživateli vypíše všechny dostupné informace o předmětu dotazu. Data bohužel není možné stáhnout bezplatně, uživatel musí zaplatit předplatné. Dostupné typy formátů pro stažení jsou rastrový obrázek, vektorový obrázek nebo ve formátu podporující programovací jazyk wolfram.

Tento zdroj poskytuje službu API (za poplatek). Základní prefix dotazu je **http://api.wolframalpha.com/v1/simple?appid=**, po kterém následuje uživateli přidělené identifikační číslo. Dále dotaz obsahuje znak **&**, a poté parametr **i**, ke kterému je přiřazen dotaz ve formě textu (mezery jsou nahrazeny znakem „+“). Následují nepovinné parametry oddělené znakem **&**. Odpovědí služby je obrázek, který by byl uživateli zobrazen při zadání hodnoty parametru **i** do webové služby tohoto zdroje. Více informací o této službě je dostupné na internetových stránkách tohoto zdroje. (Wolfram Alpha LLC, 2018)

2.1.8 Český statistický úřad

Český statistický úřad (ČSÚ) je ústředním orgánem státní správy České republiky. Byl zřízen dne 8. ledna 1969 zákonem č. 2/1969 Sb., o zřízení ministerstev a jiných ústředních orgánů státní správy. Jako svůj poradní orgán zřizuje Českou statistickou radu podle § 6 zákona č. 89/1995 Sb., o státní statistické službě. V jejím čele stojí předseda Českého statistického úřadu. Členové rady jsou jmenováni z řad odborníků statistické teorie a praxe. Rada má nejméně 11 a nejvíce 25 členů. Rada projednává program statistických zjišťování, její další úkoly a způsob práce upravuje její statut, který vydává předseda Českého statistického úřadu. Ke dni 6.3.2018 není funkce předsedy ČSÚ obsazena, jeho funkci v současné době zastupují místopředsedové Ing. Marek Rojíček, Ph.D. a Ing. Eva Krumpová. (Český statistický úřad, 2018)

Získávání dat ČSÚ

Veřejná databáze dat ČSÚ se nachází na internetové adrese https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=home. Zde jsou dostupná všechna

veřejně dostupná data poskytovaná Českým statistickým úřadem. Data mohou být zobrazena v tabulce, pomocí grafu nebo na mapě (pokud jsou data dostupná pro menší územní celky). Dále je možné zobrazit všechna dostupná data pro jednotlivá území (obce, kraje, okresy atd.). Všechna data je možné stáhnout ve formátech XLS, XML, PDF anebo PNG v případě grafů a map. Dále jsou v záložce **Metodika** popsány detaily pro každý ukazatel (kód ukazatele, dostupná časová rozmezí, definice, skupina atd.).

Služba API Českého statistického úřadu využívá dotazovací jazyk SPARQL, který je podobný dotazovacím jazyku v relačních databázích. Dotaz v tomto formátu je možné generovat na této internetové adrese ČSÚ.

2.1.9 Eurostat

Eurostat je statistickým úřadem Evropské unie. Je organizační složkou Evropské komise na úrovni generálního ředitelství. Statistické orgány byly součástí evropské integrace již od roku 1953, zkratka Eurostat se začala používat v roce 1959. Úkolem Eurostatu je předkládat harmonizovaná statistická data na úrovni celé EU a zároveň poskytovat statistické srovnání regionů a členských států. Jeho ekonomická data také slouží jako základní a oficiální podklad pro rozhodování Evropské centrální banky, a dalších unijních institucí, v ekonomických otázkách. Eurostat veškerá svoje data získává od organizací pověřených jednotlivými členskými státy ke shromažďování statistických dat na jejich území. (Wikimedia Foundation, 2018)

Získávání dat Eurostatu

Internetová adresa databáze tohoto zdroje je http://ec.europa.eu/eurostat/data/database. Zde je zobrazený obsah databáze členěný dle obsažených odvětvích a uživateli je tím pádem umožněno v přehledném seznamu najít požadovaná data a jedním kliknutím je stáhnout ve formátu TSV nebo zobrazit přímo v prohlížeči. Data jsou zobrazena v tabulce, ale uživatel si je může zobrazit vykreslené do grafu nebo v podobě mapy. U každého ukazatele jsou dostupné detaily (např. poslední aktualizace dat, nejstarší dostupná data, kód ukazatele atd.). Dále je možné data stáhnout pro vlastní potřebu ve formátech XLS, HTML, XML, PDF a TSV.

Služba API od Eurostatu používá protokol REST. Prefix pro každý dotaz je **http://ec.europa.eu/eurostat/wdds/rest/data/v2.1**. Následují argumenty pro určení formátu (JSON nebo unicode), jazyku (dostupné jsou angličtina, francouzština a němčina) a kódu požadovaného data setu. Tyto parametry jsou odděleny lomítkem.

Dále následuje otazník a parametry, které upřesňují podobu dat (např. zaokrouhlení hodnot, jednotky atd.). Více informací o službě API eurostatu je dostupné na internetové adrese tohoto zdroje. (Eurostat, 2018)

Práce s knihovnou programu R eurostat, která zpřístupňuje službu API Eurostatu pro syntaxi programovacího jazyka R, je ukázána v kapitole 3.3 této bakalářské práce.

2.1.10 Quandl

Quandl je platforma pro finanční, ekonomické a alternativní data z celého světa provozována společností Quandl, Inc., jejíž ředitelství se nachází v Torontu v Kanadě. Projekt byl spuštěn v roce 2013. Shromažďuje data z více než pět set zdrojů, mezi něž patří mimo jiné Spojené národy, Světová banka a centrální banky. Všechna data jsou dostupná prostřednictvím API a jsou k dispozici balíčky pro několik programovacích jazyků (R, Python, Matlab a další). Některá data jsou dostupná bezplatně a pro jiné je nutná platba. V současné době má přes dvě stě tisíc uživatelů. (Wikimedia Foundation, Inc., 2018)

Získávání dat Quandlu

Internetová adresa tohoto zdroje je https://www.quandl.com/search. Zde se nachází seznam databází, ze kterých může uživatel čerpat data. V květnu 2018 je počet těchto databází tři sta dvacet (z toho sto devadesát je přístupných zdarma). Každá databáze poskytuje data, která jsou uživateli ihned prezentována pomocí grafů a tabulek. Uživatel má možnost tato data měnit dle svých potřeb, například změnou časového úseku zobrazovaných dat, frekvencí dat nebo může data transformovat (například místo hodnoty záznamu se zobrazí změna oproti hodnotě předchozího záznamu). Pro stáhnutí dat pro vlastní potřebu je nutné, aby si uživatel vytvořil účet na stránkách Quandlu a přihlásil se k tomuto účtu. Možné formáty pro stažení dat jsou CSV, XML a JSON. Přihlášený uživatel má navíc možnost přidat odkazy na data do oblíbených ve svém profilu a při pozdější návštěvě tato data načíst rychleji pomocí záložky „Bookmarks“ ve svém profilu.

Každá databáze navíc poskytuje dokumentaci ke službě API poskytované Quandlem. Tato dokumentace popisuje základní strukturu dotazů, kód dané databáze a kódy pro všechny parametry, které jsou dostupné (například kódy ukazatelů, států atd.). Dále každá databáze obsahuje záložku, kde jsou ukázky dotazů pomocí služby API,

knihoven Quandl pro programovací jazyky python a R, a také pro doplněk Quandl pro software Microsoft Excel. Knihovna Quandl pro program R je popsána v kapitole 3.1 této bakalářské práce.

2.2 Programovací prostředí R

R je integrovaná sada softwarových nástrojů pro manipulaci s daty, kalkulaci a grafické zobrazení. Výraz „prostředí“ zde představuje plně plánovaný a logický systém. Kromě jiných věcí má

- efektivní zpracování dat,
- sadu operátorů pro výpočty na polích, zejména na maticích,
- obsáhlou, souvislou, integrovanou kolekci pokročilých nástrojů pro analýzu dat,
- grafické možnosti pro analýzu dat a zobrazení buď přímo v počítači, nebo na vytisknuté kopii a
- dobře vyvinutý, jednoduchý a efektivní programovací jazyk (nazývaný ´R´), který obsahuje podmínky, cykly, uživateli definované rekurzivní funkce a vstupové a výstupové možnosti. (W. N. Venables, 2016)

R je open-source software vyvíjený a podporovaný skupinou dobrovolníků z mnoha zemí. Středem těchto dobrovolníků je skupina zvaná R-core. Tento software je multiplatformní, může být instalován na Windows, Mac OS X a Linux. Základní systém je podporován více než tisíce balíčky dostupně převážně z centrální databáze tohoto projektu cran.r-project.org a z mnoha jiných. (Chambers, 2008)

Tento projekt vznikl jako výzkum Rosse Ihaka a Roberta Gentlemana v devadesátých letech, popsán byl v roce 1996. Od té doby se rozšířil natolik, že obsahuje většinu nových statistických technik. Software v R implementuje verzi programovacího jazyka S, který byl vyvinut mnohem dříve skupinou stávající z Ricka Beckera, Johna Chamberse a Allana Wilkse v Bell Laboratories a popsán v sérii knih. (Chambers, 2008)

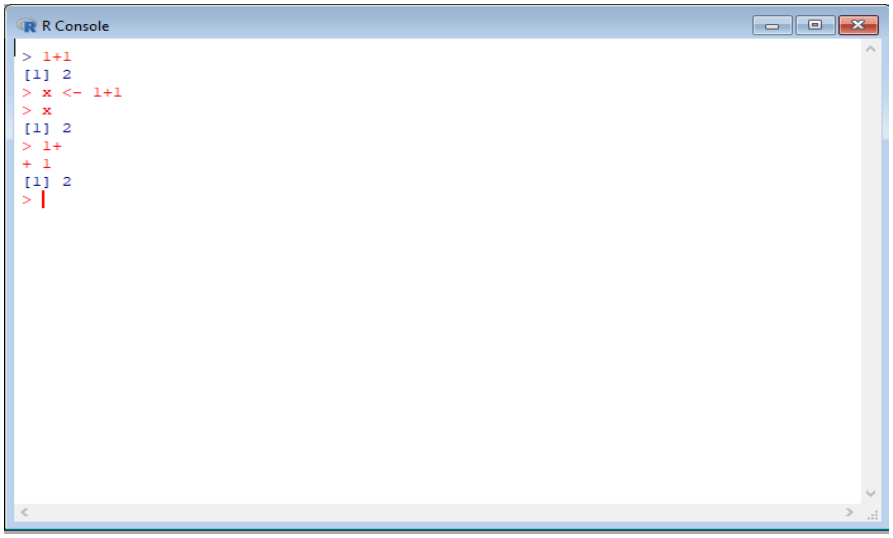
Většina programu R je napsaná ve stejném jazyce, který je používán pro komunikaci se systémem, tedy odnoží programovacího jazyka S. Tento jazyk se vyvinul do současné podoby v osmdesátých letech ve funkcionálním paradigmatu. Základní jednotkou tohoto programování je funkce. Následující vývoj jazyka představil třídy a metody. Metody jsou speciální funkce patřící určité třídě s jedním nebo více argumenty.

Třídy definují obsah objektů. R přidalo několik funkcí k tomuto jazyku, stále však zůstává z velké části kompatibilní s jazykem S. (Chambers, 2008)

Program R je dostupný na internetové adrese https://cran.r-project.org/. Uživatel má na výběr verzi pro Linux, Mac OS X a Windows.

V prostředí programu R ve Windows se pracuje podobně jako v příkazovém řádku. Ukázka této konzole je na obrázku 1. Znak „>“ značí, že program R očekává zadání příkazu od uživatele. Pro provedení příkazu je potřeba stisknout klávesu „Enter“. Dále je na obrázku 1 ukázka základní syntaxe programovacího jazyka R. Pokud uživatel zadá příkaz program mu zobrazí výsledek tohoto příkazu modrým písmem. Soubor znaků „<-“ značí přiřazení hodnoty nebo hodnot do proměnné, v tomto případě program R nevypíše výsledek příkazu, ale pouze ho uloží do paměti dané proměnné. Pro vypsání obsahu proměnné je třeba zadat název proměnné jako samostatný příkaz. Pokud uživatel zadá neúplný příkaz, program R vypíše na další řádku „+“ a uživatel je nucen pokračovat v příkazu.

Obrázek 1: Ukázka prostředí programu R



Zdroj: vlastní tvorba v programu R

2.3 API

„API (zkratka pro Application Programming Interface) označuje v informatice rozhraní pro programování aplikací. Tento termín používá softwarové inženýrství. Jde o sbírku procedur, funkcí, tříd či protokolů nějaké knihovny (ale třeba i jiného programu nebo jádra operačního systému), které může programátor využívat. API určuje, jakým způsobem jsou funkce knihovny volány ze zdrojového kódu programu. Při použití v

kontextu vývoje webu je API typicky definováno HTTP a požaduje zprávy spolu s definicí struktury odpovědí obvykle v XML nebo JSON formátu. Zatímco „Web API“ je prakticky synonymem pro Webovou službu, nedávný trend (tak zvaný Web 2.0) se vzdaluje od Simple Object Access Protocol, služby založené na více přímých REST stylů komunikace. Web API umožňují kombinaci různých služeb do nových aplikací známých jako mashups.“ (Wikimedia Foundation, Inc., 2018)

3. Praktická část

V této kapitole jsou popsány čtyři knihovny programu R, které umožňují uživateli získávat data z webových databází pomocí syntaxe programu R. U každé knihovny je popsána instalace knihovny, postup při získávání dat a praktická ukázka získání dat a základní zpracování těchto dat. Ukázky příkazů v programu R jsou znázorněny následujícím stylem.

```
> „Příkaz“
```

3.1 Knihovna Quandl

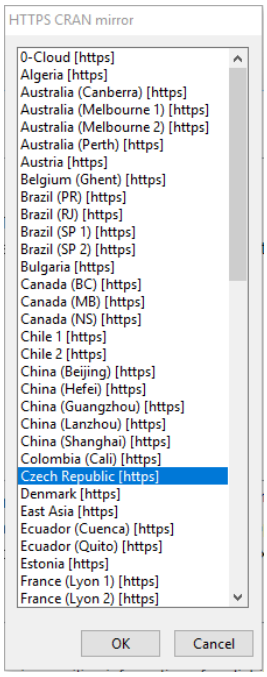
3.1.1 Instalace knihovny Quandl

Pro instalaci knihovny Quandl v programu R je pouze nutné zadat následující příkaz a provést instalaci balíčku.

```
> install.packages ("Quandl")
```

Dále je uživatel vyzván vybrat tzv. mirror (viz. obrázek 2), což je web, který kopíruje jeho originál. Knihovna by se nainstalovala výběrem jakéhokoliv mirroru, ale pro větší efektivitu a rozložení zatížení hlavního serveru je doporučeno vybírat mirro určený pro Českou republiku.

Obrázek 2: Výběr mirroru v programu R



Zdroj: vlastní tvorba v programu R

Pro použití této knihovny v daném kontextu programu slouží příkaz.

```
> library (Quandl)
```

Pokud uživatel plánuje provést více než padesát dotazů denně, je nutné zaregistrovat si účet na internetových stránkách Quandlu. Po zaregistrování je uživateli přiřazen osobní API klíč, který mu umožní provádět neomezený počet dotazů denně a také zaručuje přístup k placeným databázím, které jsou předplacené pro účet s tímto API klíčem. Pro zadání API klíče do programu R slouží tento příkaz.

```
> Quandl.api_key ("APIKLÍČ")
```

3.1.2 Získávání dat pomocí knihovny Quandl

Základní struktura příkazu pro získání dat prostřednictvím knihovny Quandl v programu R je následující

```
> Quandl (´kód databáze/kód data setu´, …)
```

Vyhledávání databází je k dispozici na této internetové adrese https://www.quandl.com/search. V dokumentaci databáze je uveden její kód. Každá databáze navíc poskytuje seznam data setů a jejich kódů, které si uživatel může stáhnout a pomocí těchto seznamů vyhledávat požadovaná data. Dále příkaz může obsahovat další parametry, které se přidají do závorky příkazu Quandl. Následují po kódech databáze a data setu, a jsou oddělené čárkou. Dostupné parametry jsou uvedené v následující tabulce.

Seznam dostupných parametrů pro příkaz **Quandl ()** je vypsán v tabulce 1**.**

Tabulka 1: Seznam parametrů pro příkaz Quandl()

Parametr	Požadován	Typ	Hodnoty	Popis
Kód databáze	ANO	textový řetězec		Kód, který určuje databázi, z které se získávají data.
Kód datasetu	ANO	textový řetězec		Kód, který určuje dataset.
limit	NE	číslo		limit=n příkaz načte prvních n řádků v datasetu. limit=1 vrátí poslední řádek.
column_index	NE	číslo		Požádá o načtení určitého sloupce. Sloupec 0 obsahuje

				datumy. Data začínají ve sloupci 1.
start_date	NE	textový řetězec	RRRR-MM-DD	Načte data od zadaného datumu a novější.
end_date	NE	textový řetězec	RRRR-MM-DD	Načte data až do zadaného datumu a starší.
order	NE	textový řetězec	asc desc	Načte data ve vzestupném nebo sestupném pořadí datumů. Výchozí desc (sestupné).
collapse	NE	textový řetězec	none daily weekly monthly querterly annual	Změní frekvenci načtených dat. Výchozí je none – data jsou načtena v originální granularitě. daily = denní, weekly = týdenní, monthly = měsíční, quarterly = čtvrtletní, annual = roční
transform	NE	textový řetězec	none diff rdiff rdiff_from cumul normalize	Provede základní výpočty na datech před načtením. Výchozí je none. Možnosti výpočtů jsou popsány v tabulce níže.
type	NE	textový řetězec	raw ts zoo xts timeSeries	Určuje v jakém formátu se data načtou.

Zdroj: https://docs.quandl.com/docs/parameters-2

Možnosti výpočtů na datech před načtením pomocí knihovny Quandl jsou vypsány v tabulce 2.

Tabulka 2: Možnosti výpočtů na datech před načtením pomocí knihovny Quandl

Název	Efekt	Vzorec
none	žádný efekt	z[t] = y[t]

diff	změna oproti předchozí hodnotě	z[t] = y[t] – y[t-1]
rdiff	změna oproti předchozí hodnotě v %	z[t] = (y[t] – y[t-1]) / y[t-1]
rdiff_from	poslední hodnota je přírůstek	z[t] = (y[poslední] – y[t]) / y[t]
cumul	kumulativní suma	z[t] = y[0] + y[1] + … + y[t]
normalize	škáluje řadu, aby začala na 100	z[t] = y[t] ÷ y[0] * 100

Zdroj: https://docs.quandl.com/docs/parameters-2

3.1.3 Ukázka práce s knihovnou Quandl

V této praktické ukázce knihovny Quandl je ukázáno získání hodnot hrubého domácího produktu pro Českou republiku, Slovensko a Rumunsko. Následně jsou tyto hodnoty vykresleny do spojnicového grafu pomocí příkazů programu R.

Spojnicový graf se v programu R vytváří pomocí funkce **plot()**, která jako první dva argumenty přijímá vektory se souřadnicemi na ose x a y nebo jednu proměnnou (strukturu, funkci nebo jiný objekt programu R), která má metodu pro vykreslení. Tímto grafem se nejlépe zachycuje vývoj určitých dat v čase. Souřadnice x tedy bude znázorňovat čas, v našem případě jednotlivé roky, a souřadnice y bude obsahovat hodnoty ukazatele pro jednotlivé státy.

Nejdříve vykreslení hrubého domácího produkt pro Českou republiku. Do nové proměnné jsou pomocí knihovny Quandl z databáze IMF načtena data týkající se hrubého domácího produktu České republiky. **ODA** je název databáze, **CZE** kód státu a **NGDPD** je kód data setu:

```
> hdpcz <- Quandl ('ODA/CZE_NGDPD')
```

Další argument pro funkci **plot()** je **type**, který určuje, jak bude výsledný graf vykreslen. K tomuto argumentu se přiřazují písmena nadefinovaná v manuálu pro R. Příklady těchto písmen jsou **p** pro vykreslení bodů, **l** pro vykreslení čar, **o** pro vykreslení bodů i čar. Na každém uživateli je zvolit si vlastní styl, zde je pro tento argument použito písmeno **o**. Následuje argument **col** určující barvu čar nebo bodů v grafu. Argument **ylim** určuje limitní hodnoty na ose y (obdobně se může určit limit na

ose x argumentem **xlim**). Funkce **max()** vrátí uživateli nejvyšší číselnou hodnotu v daném vektoru. Výsledný příkaz tedy vypadá následovně:

```
> plot (hdpcz, type = "o", xlab = "Rok", ylab = "Hodnota HDP v miliardách dolarů", col = 'blue', ylim = c (0, max (hdpcz$Value)))
```

Nyní je vykreslen spojnicový graf znázorňující hodnotu hrubého domácího produktu v miliardách dolarů v letech 1995 až 2022. Dále jsou přidány další státy, například Slovensko a Rumunsko. Nejdříve, podobně jako u České republiky, vytvoření nových proměnných a načtení hodnot hrubého domácího produktu daného státu.

```
> hdpsvk <- Quandl ('ODA/SVK_NGDPD')
> hdprou <- Quandl ('ODA/ROU_NGDPD')
```

Pro vykreslení hodnot dalšího státu do grafu se používá funkce **lines()**. Tato funkce bude mít tři argumenty, které jsou definované už u funkce **plot()**, a to data frame s příslušnými daty, **type** a **col**. Typ vykreslení zůstává stejný jako v prvním případě a barva je změněna pro každý stát.

```
> lines (hdpsvk, type = "o", col = "red")
> lines (hdprou, type = "o", col = "green")
```

Každý správný graf by měl obsahovat legendu a nadpis. Pro vytvoření legendy se používá funkce **legend()**. Argumenty této funkce jsou souřadnice x a y určující, kde se bude legenda nacházet, nebo slovní určení místa legendy v angličtině (možné hodnoty jsou k nalezení v manuálu), vektor obsahující popisky pro každou vykreslenou čáru, **col** v podobě vektoru, kde budou vypsány barvy čar ve stejném pořadí jako popisky a **lty** určující typ čar vykreslených v legendě.

```
> legend („topleft“, c("Česka republika", "Slovensko", "Rumunsko"), col = c ("blue", "red", "green"), lty = 1)
```

Pro vypsání názvu grafu se používá funkce **title()**. Zde jsou použity argumenty **main** určující hlavní název grafu a **font.main**, kterým se upraví font názvu.

```
> title (main = "HDP v České republice, Slovensku a Rumunsku v letech 1995 až 2022", font.main = 4)
```

Pro uložení grafu do souboru se může použít horní menu, kde je v záložce **File -> Save as** možnost vybrat formát, ve kterém chceme graf uložit. Na výběr jsou zde

například formáty PDF, JPG, PNG, BMP. Druhá možnost je uložit graf pomocí příkazů. První z příkazů nutný pro uložení grafu je funkce nazvaná podle požadovaného formátu. Jako argument se této funkci předává název souboru, do kterého chce uživatel graf uložit. Pokud chce uživatel souboru uložit jinam než do pracovního adresáře, musí zadat úplnou cestu k souboru (absolutní nebo relativní). Následují příklady uložení v různých formátech:

```
> pdf ("D:\škola\Bakalářka\Obrázky\hdp-spoj-cz,sk,cu.pdf")
```

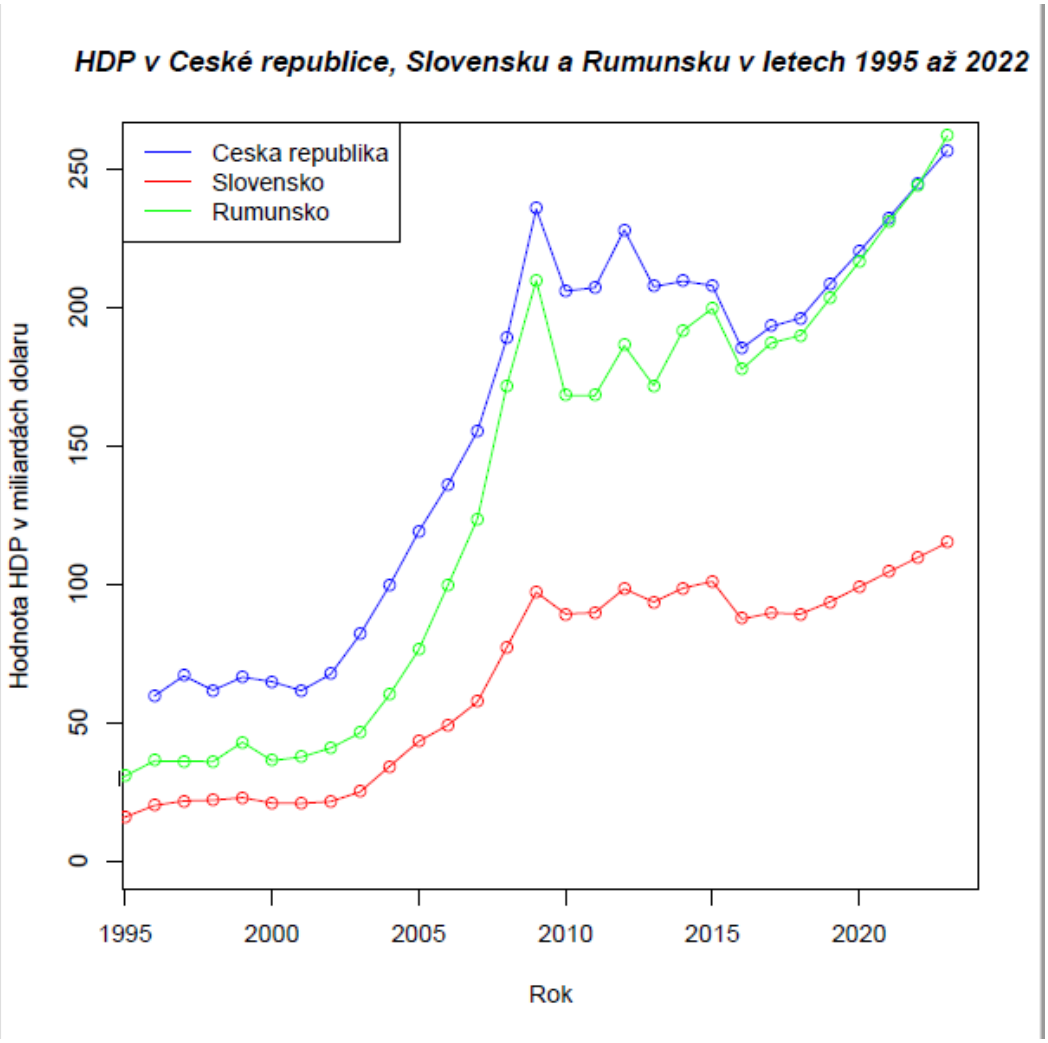
```
> jpeg ("hdp-spoj-cz,sk,cu.jpg")
```

Po tomto příkazu je potřeba zopakovat příkazy **plot()**, **lines()**, **legend()** a **title()** se všemi výše uvedenými argumenty. Po těchto příkazech se vykreslování do souboru vypíná následujícím příkazem:

```
> dev.off ()
```

Výsledný graf je k vidění na obrázku 3.

Obrázek 3: Spojnicový graf



Zdroj: vlastní tvorba v programu R

Pro lepší porovnání vývoje hrubého domácího produktu jednotlivých států lze data přetransformovat do podoby, kde hodnota v daném roce odpovídá procentní změně tohoto ukazatele oproti roku předchozímu. Tuto transformace je možné provést již při získávání dat z databází Quandl pomocí argumentu **transofrm**, který je popsán v tabulce v kapitole 4.1.2 této práce, nebo přímo v programu R.

Pro transformaci dat v programu R je možné využít nadefinované funkce, které jsou součástí některých knihoven, které pracují s data framy nebo časovými řadami, nebo cyklus **for**, který se používá ve většině programovacích jazyků. Příklad takového cyklu, který využívá proměnnou **hdprou** z předchozího příkladu, vypadá následovně:

```
> for (i in 1 : length(hdprou$Value)) {
+ hdprou$Value[i] = (hdprou$Value[i] - hdprou$Value[i+1]) / hdprou$Value[i+1]}
```

Proměnná **i** se v cyklech při programování standardně používá pro označení indexu, který určuje pořadí prvku ve vektoru či jiné struktuře (v programu R je první index 1, v jiných jazycích může být první index označen číslem 0). Argument **1:length(hdprou$Value)** říká, že proměnná **i** bude v první iteraci cyklu 1, a v každé další iteraci se zvětší o jedna, dokud bude menší nebo rovná počtu prvků ve vektoru **hdprou$Value** (počet prvků v proměnné zjišťuje funkce **length()**). V proměnné **hdprou** pracujeme pouze se sloupcem **Value**. Ve složených závorkách se nachází operace, která proběhne v každé iteraci cyklu právě jednou. Je zde použit vzoreček **z[t] = (y[t] – y[t+1]) / y[t+1]**, jelikož v proměnné **hdprou** jsou hodnoty seřazené od nejnovější po nejstarší.

V případě transformace dat již při získávání dat by příkazy vypadaly následovně:

```
> hdprou <- Quandl ('ODA/ROU_NGDPD', start_date = "1995-12-31", end_date = "2017-12-31", transform = "rdiff")
> hdpcz <- Quandl ('ODA/CZK_NGDPD', start_date = "1995-12-31", end_date = "2017-12-31", transform = "rdiff")
> hdpsvk <- Quandl ('ODA/SVK_NGDPD', start_date = "1995-12-31", end_date = "2017-12-31", transform = "rdiff")
```

Nyní převedení hodnot do nových proměnných a obrácení pořadí prvků pomocí funkce **rev()**:

```
> cz <- rev (hdpcz$Value)
> svk <- rev (hdpsvk$Value)
> rou <- rev (hdprou$Value)
```

Funkce **summary()** vypíše uživateli základní statistické informace o vektoru, který je předán jako argument.

Obrázek 4: Ukázka funkce summary()

```
> summary(cz)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-12.590  -2.029   5.967   6.178  14.260  24.870
> summary(svk)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-13.4800  -0.1882   3.7650   7.8280  15.6000  36.8900
> summary(rou)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-19.8500  -0.1097   7.0190   8.9280  21.4300  38.9700
```

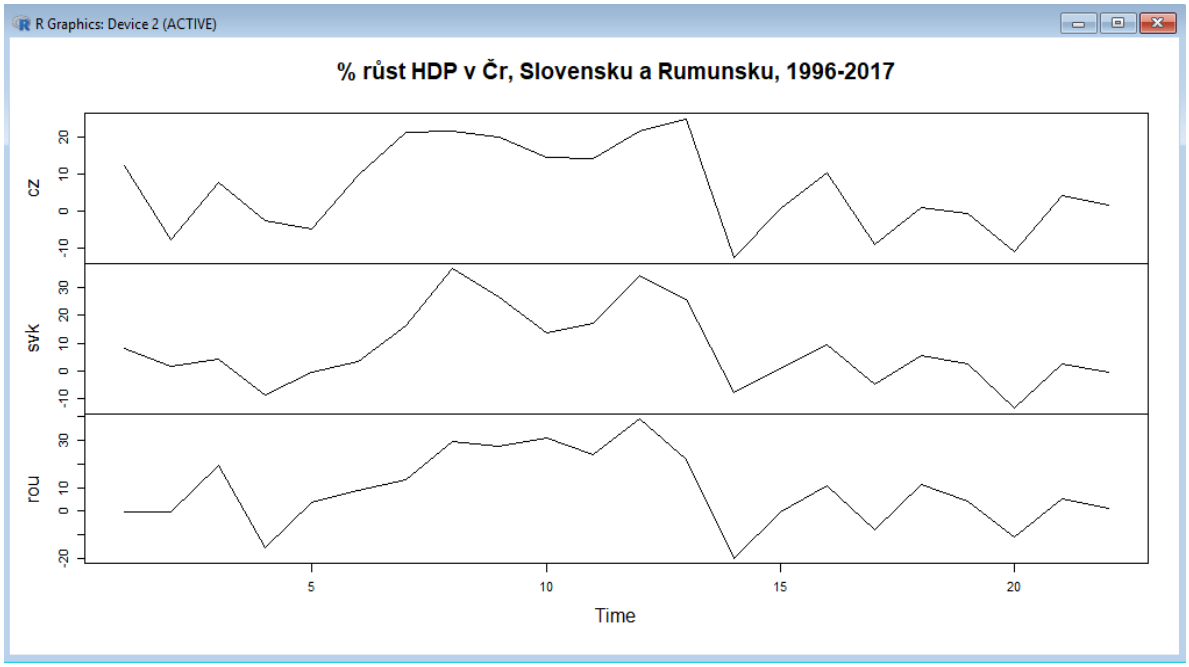
Zdroj: vlastní tvorba v programu R 1

Nakonec sloučení vektorů do jednoho data framu a vykreslení pomocí funkce **plot.ts():**

```
> hdp <- cbind (cz, svk, rou)
> plot.ts (hdp)
```

Výsledné vykreslení časových řad je ukázáno na obrázku 5.

Obrázek 5: Vykreslení časových řad



Zdroj: vlastní tvorba v programu R

3.2 Knihovna wbstats

3.2.1 Instalace knihovny wbstats

Pro instalaci knihovny wbstats v programu R je nutné zadat následující příkaz a provést instalaci balíčku.

```
> install.packages ("wbstats")
```

Dále je uživatel vyzván vybrat tzv. mirror (viz. obrázek 2), což je web, který kopíruje jeho originál. Knihovna by se nainstalovala výběrem jakéhokoliv mirroru, ale pro větší efektivitu a rozložení zatížení hlavního serveru je doporučeno vybírat mirro určený pro Českou republiku.

Pro použití této knihovny v daném kontextu programu R slouží příkaz:

```
> library (wbstats)
```

3.2.2 Získávání dat pomocí knihovny wbstats

Pokud uživatel neví kód státu a indikátoru, který chce získat, první krok je získání těchto kódů pomocí následujícího příkazu:

```
> wbsearch ()
```

Tento příkaz vrátí uživateli jako odpověď seznam všech dostupných indikátorů, které jsou dostupné pomocí API Světové banky, a informace o nich.
V tomto příkazu mohou být použity tři argumenty. První je **pattern**, kde je předán textový řetězec, a příkaz vrátí pouze informace o indikátorech, které obsahují daný text. Dodatek k tomuto argumentu je další argument **fields**, kde uživatel zadá název sloupce, ve kterém by měl být vyhledán text z argumentu **pattern**. Praktické využití může být například, pokud uživatel hledá data z jednoho konkrétního zdroje. Argument **extra = TRUE** slouží ke zobrazení všech informací o indikátorech. Pokud je tento argument nastaven na FALSE (primární nastavení), program vrátí pouze ID a název indikátoru.

Získání kódů států a regionů je možné pomocí data framu **countries**, který je obsažen v proměnné **wb_cachelist**, která je předdefinovaná v této knihovně. Jak je ukázáno na obrázku 7, tato proměnná obsahuje celkem sedm data framů, které dávají uživateli informace o státech, indikátorech, zdrojích dat, katalozích dat (v podstatě detailnější informace o zdrojích dat), tématech dat (ekonomická, sociální atd.), úrovni příjmů a o poskytovatelích úvěrů v rámci světové banky (IDA – Mezinárodní asociace pro rozvoj, IBRD – Mezinárodní banka pro obnovu a rozvoj).

Obrázek 6: Struktura proměnné wb_cachelist()

```
> str(wb_cachelist, max.level = 1)
List of 7
 $ countries  :'data.frame':    304 obs. of  18 variables:
 $ indicators :'data.frame':    16978 obs. of  7 variables:
 $ sources    :'data.frame':    43 obs. of  8 variables:
 $ datacatalog:'data.frame':    238 obs. of  29 variables:
 $ topics     :'data.frame':    21 obs. of  3 variables:
 $ income     :'data.frame':    7 obs. of  3 variables:
 $ lending    :'data.frame':    4 obs. of  3 variables:
```

Zdroj: vlastní tvorba v programu R

Pro získání nejnovějších informací ohledně API Světové banky slouží následující příkaz:

> wb_cachelist <- wbcache (lang = „en“)

Pro tento příkaz může být použit argument **lang**, který specifikuje použitý jazyk. Dostupné jazyky jsou angličtina, španělština, francouzština, arabština a mandarínština. Pokud uživatel nezadá tento argument, je primárně nastaven jazyk angličtina.

Pro samotné získání dat slouží příkaz:

> wb (kód státu, indikátor, …)

Seznam dostupných parametrů pro příkaz **wb()** je zobrazen v tabulce 3.

Tabulka 3: Seznam dostupných parametrů pro příkaz wb()

Parametr	Požadován	Typ	Hodnoty	Popis
country	NE	textový řetězec		Kód, který určuje stát či oblast. Primárně je nastaven na „all“.
indicator	ANO	textový řetězec		Kód, který určuje indikátor.
mrv	NE	číslo		Příkaz načte nejnovějších n hodnot v data setu. 1 vrátí poslední hodnotu.
return_wide	NE	logická hodnota	TRUE FALSE	Pokud je nastaven na TRUE, příkaz načte data ve změněném formátu. Primární nastavení je FALSE

startdate	NE	číslo, textový řetězec	2000 2000M01 2000Q1	Načte data od zadaného datumu a novější. M – měsíční frekvence, Q – čtvrtletní frekvence
enddate	NE	číslo, textový řetězec	2017 2017M01 2017Q1	Načte data až do zadaného datumu a starší. M – měsíční frekvence, Q – čtvrtletní frekvence
gapfill	NE	logická hodnota	TRUE FALSE	Zaplňuje chybějící hodnoty, na základě přechozích hodnot.
freq	NE	textový řetězec	Q M Y	Určí frekvenci načtených dat. M = měsíční, Q = čtvrtletní, Y = roční
cache	NE	list	list data framů z wbcache	Pokud argument není zadán, je použito wb_cachelist.
removeNA	NE	logická hodnota	TRUE FALSE	Pokud je nastaven na TRUE, vymaže chybějící hodnoty.

Zdroj: https://www.rdocumentation.org/packages/wbstats/versions/0.2/topics/wb

3.2.3 Ukázka práce s knihovnou wbstats

V této podkapitole je ukázka získání dat týkající se poměru pracujících a studujících lidí ve věku devatenáct až dvacet čtyři let ve čtyřech státech pomocí knihovny wbstats. Tato data jsou dále vykreslena do koláčového (nebo také kruhového) grafu.

Příkazy pro získání dat pomocí příkazu **wb():**

```
> employed <- wb (indicator="4.0.work.19a24", startdate = 2014, enddate = 2014)
> employedschool <- wb (indicator="4.0.studwork.19a24", startdate = 2014, enddate = 2014)
> school <- wb (indicator="4.0.stud.19a24", startdate = 2014, enddate = 2014)
> none <- wb (indicator="4.0.nini.19a24", startdate = 2014, enddate = 2014)
```

Zde jsou vytvořeny čtyři proměnné. **Employed** obsahuje data o pracujících lidech, **employedschool** o pracujících a studujících, **school** o studujících a **none** pro nepracující a nestudující obyvatele ve věku 19 až 24 let. Proměnné obsahují data pro všechny dostupné země pro daný rok. Struktura proměnných je vidět na obrázku 7.

Obrázek 7: Struktura proměnné employed

```
> str(employed)
'data.frame':   19 obs. of  7 variables:
 $ iso3c      : chr  NA NA NA NA ...
 $ date       : chr  "2014" "2014" "2014" "2014" ...
 $ value      : num  0.47 0.347 0.43 0.5 0.474 ...
 $ indicatorID: chr  "4.0.work.19a24" "4.0.work.19a24" "4.0.work.19a24" "4.0.wo$
 $ indicator  : chr  "Youth: Employed (19-24)" "Youth: Employed (19-24)" "Youth$
 $ iso2c      : chr  NA NA NA NA ...
 $ country    : chr  "Andean Region" "Argentina" "Bolivia" "Brazil" ...
> 
```

Zdroj: vlastní tvorba v programu R

Hodnoty všech čtyř ukazatelů jsou zde převedeny do jedné proměnné pro jednotlivé státy:

```
> brazilie <- c (employed[4, 'value'], employedschool[4, 'value'], school[4, 'value'], none[4, 'value'])
```

```
> argentina <- c (employed[2,'value'], employedschool[2,'value'], school[2, 'value'], none[2, 'value'])
```

```
> kolumbie <- c (employed[7, 'value'], employedschool[7, 'value'], school[7, 'value'], none[7, 'value'])
```

```
> mexico <- c (employed[15, 'value'], employedschool[15, 'value'], school[15, 'value'], none[15, 'value'])
```

Z každé proměnné je vzata hodnota ukazatele, která se nachází ve sloupci **value**, a každý řádek v proměnných odpovídá stejnému státu, tudíž čísla indexů řádků při přiřazování hodnot musí být u každé proměnné stejné.

Dále vytvoření vektoru obsahujícího popisky dat pro výsledný graf. Pořadí popisků musí být stejné jako pořadí ukazatelů v proměnných, kde jsou uloženy data pro jednotlivé státy.

```
> popisky <- c ("Zaměstnaní", "Zaměstnaní a studenti", "Studenti", "Neaktivní")
```

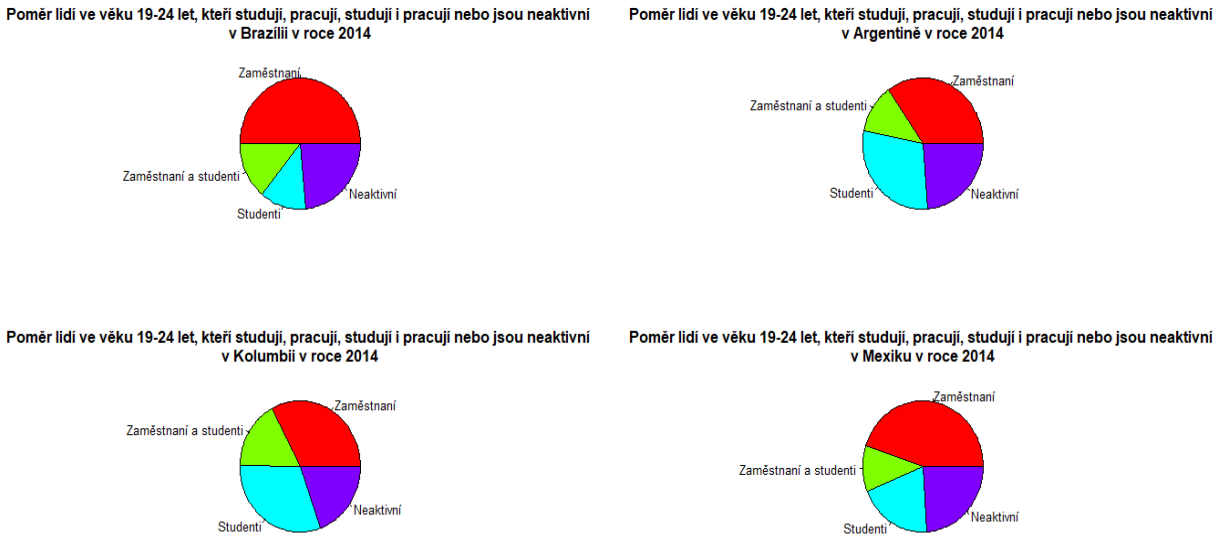
A nakonec vykreslení grafů pro každý stát zvlášť:

```
> par (mfrow = c (2,2))
> pie (brazilie, labels = popisky, col = rainbow (4), main = "Poměr lidí ve věku 19-24 let, kteří studují, pracují, studují i pracují nebo jsou neaktivní \nv Brazílii v roce 2014")
> pie (argentina, labels = popisky, col = rainbow (4), main = "Poměr lidí ve věku 19-24 let, kteří studují, pracují, studují i pracují nebo jsou neaktivní \nv Argentině v roce 2014")
> pie (kolumbie, labels = popisky, col = rainbow (4), main = "Poměr lidí ve věku 19-24 let, kteří studují, pracují, studují i pracují nebo jsou neaktivní \nv Kolumbii v roce 2014")
> pie (mexico, labels = popisky, col = rainbow (4), main = "Poměr lidí ve věku 19-24 let, kteří studují, pracují, studují i pracují nebo jsou neaktivní \nv Mexiku v roce 2014")
```

K vykreslení výsečových grafů se používá příkaz **pie().** Prvním argumentem je zde vektor vykreslovaných hodnot. Hodnoty ukazatelů mohou být v libovolných hodnotách, program R vykreslí do grafu výseče, které představují poměrné zastoupení jednotlivých hodnot z jejich součtu. Argument **labels** určuje vektor popisek jednotlivých výsečí. **Col** určuje barvy použité v grafu. Příkaz **rainbow(n)** určí, že v grafu se použije **n** různých barev (pokud má **n** menší hodnotu než počet výsečí, barvy se začnou opakovat). Poslední argument **main** určuje hlavní název grafu. V tomto případě text obsahuje výraz **\n**, který značí novou řádku v textu.

Výsledné grafy jsou k vidění na obrázku 8.

Obrázek 8: Výsečové grafy



Zdroj: vlastní tvorba v programu R

3.3 Knihovna eurostat

3.3.1 Instalace knihovny eurostat

Pro instalaci knihovny eurostat v programu R je pouze nutné zadat následující příkaz a provést instalaci balíčku.

```
> install.packages ("eurostat")
```

Dále je uživatel vyzván vybrat tzv. mirror (viz. obrázek 2), což je web, který kopíruje jeho originál. Knihovna by se nainstalovala výběrem jakéhokoliv mirroru, ale pro větší efektivitu a rozložení zatížení hlavního serveru je doporučeno vybírat mirror určený pro Českou republiku.

Pro použití této knihovny v daném kontextu programu R slouží příkaz:

```
> library (eurostat)
```

3.3.2 Získávání dat pomocí knihovny eurostat

Pro získání kódů států slouží následující příkazy:

```
> eu_countries
```

```
> ea_countries
```

> efta_countries

> eu_candidate_countries

Každý z těchto příkazů souží pro získání kódů pro rozdílnou skupinu států (**eu_countries** pro státy Evropské unie, **ea_countries** pro státy eurozóny, **efta_countries** pro členy Evropského sdružení volného obchodu a **eu_candidate_countries** pro kandidáty na členství v Evropské unii). Odpovědí programu je data frame, který obsahuje kód státu a název státu.

Získání kódů ukazatelů je možné pomocí následujícího příkazu:

> search_eurostat („text“)

Jediný povinný argument je uvozený text, který udává textový řetězec, který je obsažen v názvech ukazatelů v odpovědi programu (např. unemployment, education atd.). Dalším nepovinným argumentem je **type**, který určuje typ proměnné, ve které jsou uloženy data o daném ukazateli. Primárně je tento argument nastaven na data set, další možné hodnoty jsou **folder**, **table** nebo **all**. Poslední možný argument u tohoto příkazu je **fixed**, který má logickou hodnotu (**TRUE** nebo **FALSE**). Pokud má hodnotu **TRUE** (primárně), příkaz vyhledá přesný text z prvního argumentu, ale pokud má hodnotu **FALSE**, text může obsahovat regulární výrazy.

Pro samotné získání dat se používá následující příkaz:

> get_eurostat („id“, …)

Seznam dostupných parametrů pro příkaz **get_eurostat()** je vypsán v tabulce 4.

Tabulka 4: eznam dostupných parametrů pro příkaz get_eurostat()

Parametr	Požadován	Typ	Hodnoty	Popis
id	ANO	textový řetězec		Kód, který určuje indikátor.
time_format	NE	textový řetězec	date date_last num raw	Textový řetězec, který určuje typ proměnné, na který se převede sloupec „time“. Primární nastavení je „date“ date, date_last – Date, num – číslo, raw – žádný převod

filters	NE	list		Filtruje získaná data podle vektorů v listu, a v nich obsažených hodnot.
type	NE	textový řetězec	code label	Určuje typ proměnných. Primárně nastaven na „code“.
select_time	NE	textový řetězec	Y - roční S – půlroční Q – čtvrtletní M – měsíční	Určuje frekvenci dat.
cache	NE	logická hodnota	TRUE FALSE	Určuje, zda bude program ukládat dotazy do paměti.
update_cache	NE	logická hodnota	TRUE FALSE	Určuje, zda program před dotazem aktualizuje cache.
cache_dir	NE	textový řetězec		Určuje cestu k adresáři pro cache.
compress_file	NE	logická hodnota	TRUE FALSE	Určuje, zda bude RDS soubor komprimovaný. Primárně je nastavena hodnota TRUE.
stringAsFactors	NE	logická hodnota	TRUE FALSE	Pokud nastaven na TRUE, proměnné jsou převedeny na typ „factor“. FALSE – převedení na textový řetězec.

Zdroj: https://cran.r-project.org/web/packages/eurostat/eurostat.pdf

3.3.3 Ukázka práce s knihovnou eurostat

V této podkapitole je ukázáno získání dat o nezaměstnanosti třiceti jedna států Evropy k 1. 1. 2016, a dále manipulace a úprava těchto dat a následné vykreslení do sloupcového grafu.

Pro získání dat je použit následující příkaz:

```
> nezm <- get_eurostat (id="tps00203", filters = list (geo = c("CZ", "DE", "BE", "BG", "DK", "EE", "IE", "EL", "ES", "FR", "HR", "IT", "CY", "LV", "LT", "LU", "HU", "MT", "NL", "AT", "PL", "PT", "RO", "SI", "SK", "FI", "SE", "UK", "IS", "NO", "TR"), time = 2017))
```

K argumentu **id** je přiřazena hodnota **tps00203**, což je kód pro celkovou nezaměstnanost. Dále jsou zde použity filtry pro **geo**, kde jsou uvedeny kódy států, a **time** ve formátu RRRR. Zobrazení struktury proměnné pomocí příkazu **str()** ukáže, že odpověď programu R uložená v proměnné **nezm** obsahuje devadesát tři řádků dat (viz. Obrázek 9). Důvodem toho je, že knihovna eurostat vrátila nezaměstnanost ve třech různých jednotkách (na obrázku proměnná **unit**). První je **PC_ACT**, což je procentní nezaměstnanost ekonomicky aktivního obyvatelstva, další je **PC_POP**, které udává procentní nezaměstnanost z celkové populace státu a poslední je **THS_PER**, což je počet nezaměstnaných v tisících. Proměnná **age** a **sex** má zde pouze jednu hodnotu, protože v tomto ukazateli není sledována nezaměstnanost podle věku a pohlaví.

Obrázek 9: Struktura proměnné nezm

```
> str(nezm)
Classes 'tbl_df', 'tbl' and 'data.frame':       93 obs. of  6 variables:
 $ age   : Factor w/ 1 level "TOTAL": 1 1 1 1 1 1 1 1 1 1 ...
 $ unit  : Factor w/ 3 levels "PC_ACT","PC_POP",..: 1 1 1 1 1 1 1 1 1 1 $
 $ sex   : Factor w/ 1 level "T": 1 1 1 1 1 1 1 1 1 1 ...
 $ geo   : Factor w/ 31 levels "AT","BE","BG",..: 1 2 3 4 5 6 7 8 9 10 .$
 $ time  : Date, format: "2017-01-01" "2017-01-01" ...
 $ values: num  5.5 7.1 6.2 11.1 2.9 3.8 5.7 5.8 21.5 17.2 ...
>
```

Zdroj: vlastní tvorba v programu R

Aby bylo možné pracovat s daty pro jednotlivé jednotky ukazatele, je potřeba roztřídit proměnnou **nezm** do třech různých proměnných podle hodnoty proměnné **unit**. Toho uživatel v programu R dosáhne následujícími příkazy:

```
> nezm_pc_act <- nezm[nezm$unit == "PC_ACT",]
```

```
> nezm_pc_pop <- nezm[nezm$unit == "PC_POP",]
```

```
> nezm_ths_per <- nezm[nezm$unit == "THS_PER ",]
```

Pro vykreslení dat z jednotlivých proměnných do dvou sloupcových grafů se mohou použít následující příkazy.

```
> par (mfrow = 2:1)
```

```
> barplot (nezm_pc_act$values, names.arg = nezm_pc_act$geo, ylim =
c(0,25))
```

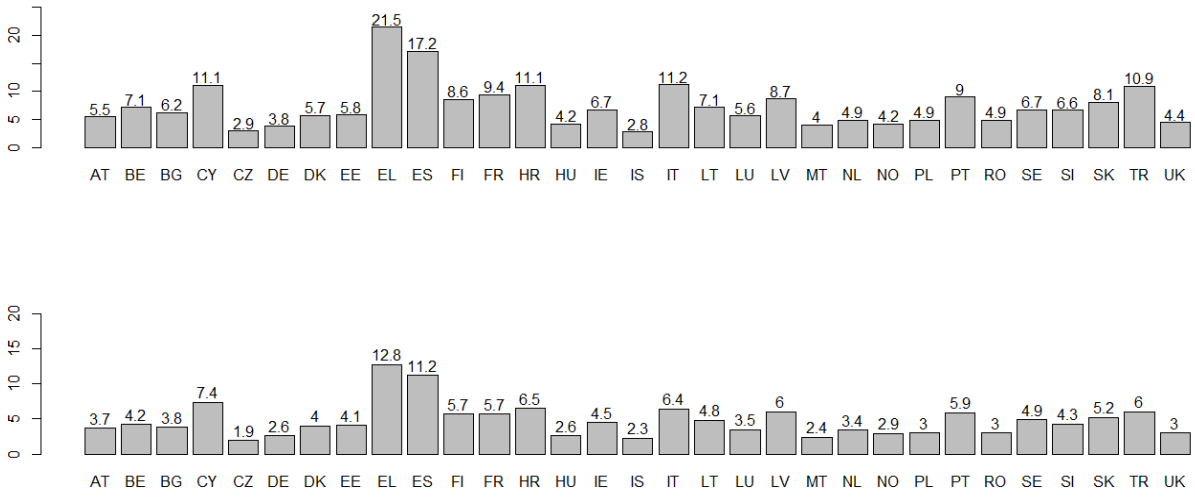
```
> text (x = seq(0.7,37,1.2), y = nezm_pc_act$values +1.5, labels = as.character
(nezm_pc_act$values))
```

```
> barplot (nezm_pc_pop$values, names.arg = nezm_pc_pop$geo, ylim = c (0,20))
> text (x = seq(0.7, 37, 1.2), y = nezm_pc_pop$values +1, labels = as.character (nezm_pc_pop$values))
```

Příkaz **par(mfrow = 2:1)** určuje že se v programu R vykreslí dva grafy pod sebou (první číslo určuje počet řádků a druhé počet sloupců). Další příkaz **barplot()** slouží k vykreslení sloupcového grafu. Povinným argumentem je vektor čísel, který určuje výšky sloupců, v tomto případě jsou to hodnoty jednotlivých ukazatelů. Nepovinnými argumenty jsou **names.arg**, který určuje popisky jednotlivých sloupců (zde kódy států) a **ylim**, který určuje rozsah hodnot zobrazovaný na ose y. Další příkaz **text()** vykresluje do grafu hodnoty jednotlivých sloupců. Jeho povinnými argumenty jsou souřadnice x a y. Jelikož v tomto případě nemá osa x číselné popisky, je pro souřadnice na ose x použita funkce **seq(a, b, c),** která vytváří posloupnost čísel (**a** = počátek posloupnosti, **b** = konec posloupnosti, **c** = diference mezi prvky). Souřadnice y je pak hodnota ukazatele (v grafu výška sloupce) plus 1,5, čímž se docílí, že hodnota je vypsána nad příslušný sloupec. Argument **labels** pak určuje text, který je vypsán na příslušné souřadnice, v tomto případě znovu hodnoty ukazatele.

Výsledné grafy jsou ukázané na obrázku 10.

Obrázek 10: Sloupcové grafy



Zdroj: vlastní tvorba v programu R

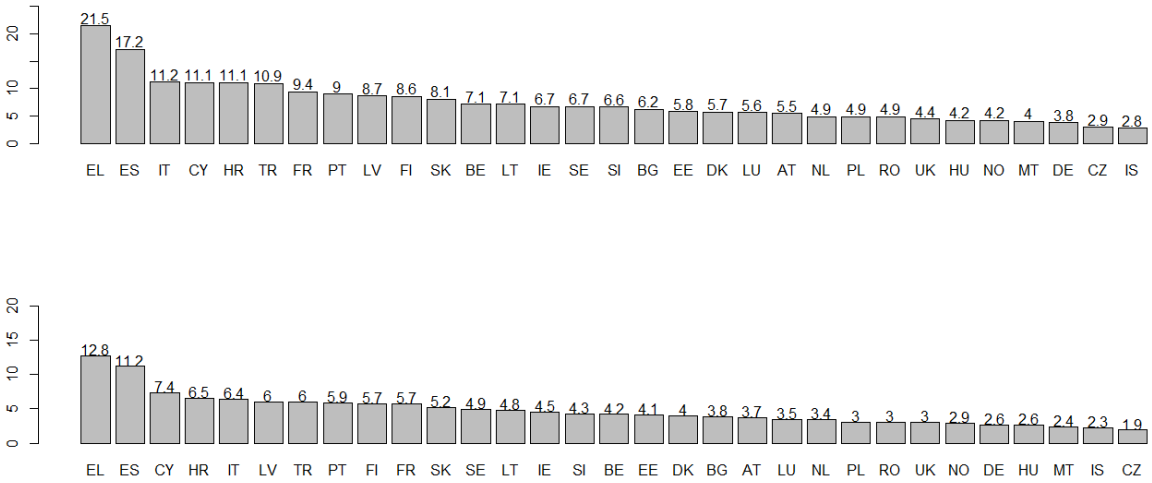
Pro lepší přehlednost je možné nejdříve záznamy v jednotlivých proměnných seřadit podle hodnot ukazatelů. Toho uživatel docílí následujícím příkazem:

```
> nezm_pc_act <- nezm_pc_act [order (nezm_pc_act$values, decreasing = TRUE),]
```

```
> nezm_pc_pop <- nezm_pc_pop [order (nezm_pc_pop$values, decreasing = TRUE),]
```

Výsledné grafy vytvořené pomocí stejných příkazů jako v předchozím případě jsou ukázány na obrázku 11.

Obrázek 11: Sloupcové grafy se setříděnými hodnotami



Zdroj: vlastní tvorba v programu R

3.4 Knihovna WDI

3.4.1 Instalace knihovny WDI

Pro instalaci knihovny WDI v programu R je pouze nutné zadat následující příkaz a provést instalaci balíčku.

```
> install.packages ("WDI")
```

Dále je uživatel vyzván vybrat tzv. mirror (viz. obrázek 2), což je web, který kopíruje jeho originál. Knihovna by se nainstalovala výběrem jakéhokoliv mirroru, ale pro větší efektivitu a rozložení zatížení hlavního serveru je doporučeno vybírat mirror určený pro Českou republiku.

Pro použití této knihovny v daném kontextu programu R slouží příkaz:

```
> library (WDI)
```

3.4.2 Získávání dat pomocí knihovny WDI

Pro získání kódů států a ukazatelů slouží v programu R následující příkaz:

```
> WDIcache ()
```

Příkazu není možné předat žádné argumenty. Odpověď programu obsahuje dvě proměnné **series** (informace o ukazatelích) a **countries** (informace o státech), které jsou typu list. Následující příkazy uloží tyto listy do nových proměnných a převedou jejich typ na data frame:

```
> series <- as.data.frame (WDIcache()$series)
```

```
> country <- as.data.frame (WDIcache()$country)
```

Struktura těchto data framů je vidět na obrázku 12. Proměnná **series** obsahuje kód ukazatele, název ukazatele, popis ukazatele, databázi, ze které pochází data, a název organizace, která tato data publikovala. Proměnná **country** obsahuje iso3 a iso2 kód státu, název státu, region, hlavní město, zeměpisnou šířku a výšku, úroveň příjmů a informace o poskytovatelích úvěrů v rámci světové banky.

Obrázek 12: Struktura proměnných series a country

```
> str(series)
'data.frame':   16695 obs. of  5 variables:
 $ indicator         : Factor w/ 16692 levels "1.0.HCount.1.90usd",..: 8206 820$
 $ name              : Factor w/ 16176 levels " Baccalaureate in Central Africa$
 $ description       : Factor w/ 5754 levels "","  Percentage of firms expected $
 $ sourceDatabase    : Factor w/ 38 levels "Africa Development Indicators",..: $
 $ sourceOrganization: Factor w/ 486 levels "","  Source: Central Bureau of  Hea$
> str(country)
'data.frame':   304 obs. of  9 variables:
 $ iso3c    : Factor w/ 304 levels "ABW","AFG","AFR",..: 1 2 3 4 5 6 7 8 9 10 .$
 $ iso2c    : Factor w/ 304 levels "1A","1W","4E",..: 25 16 13 20 18 14 147 1 1$
 $ country  : Factor w/ 304 levels "Afghanistan",..: 13 1 2 8 3 7 6 10 290 11 .$
 $ region   : Factor w/ 8 levels "Aggregates","East Asia & Pacific",..: 4 7 1 8$
 $ capital  : Factor w/ 212 levels "","Abu Dhabi",..: 134 80 1 100 191 10 1 1 2$
 $ longitude: Factor w/ 212 levels "","-0.126236",..: 47 195 1 94 126 70 1 1 18$
 $ latitude : Factor w/ 212 levels "","-0.229498",..: 58 122 1 42 146 151 1 1 1$
 $ income   : Factor w/ 5 levels "Aggregates","High income",..: 2 3 1 4 5 2 1 1$
 $ lending  : Factor w/ 5 levels "Aggregates","Blend",..: 5 4 1 3 3 5 1 1 5 3 .$
```

Zdroj: vlastní tvorba v programu R

Informace o ukazatelích je také možné vyhledat pomocí následujícího příkazu:

```
> WDIsearch (string = „text“)
```

Tento příkaz může obsahovat čtyři argumenty. Prvním je **string**, který má hodnotu ve formě textu. Tento text je vyhledán v databázi WDI a program uživateli vrátí pouze ukazatele, které obsahují tento text. Argument **field** určuje, kde bude příkaz

hledat příslušný text z argumentu **string**. Možnými hodnotami zde jsou **indicator** pro vyhledávání mezi kódy ukazatelů, **name** pro vyhledávání v názvech ukazatelů, **description** pro vyhledávání v popisech ukazatelů, **sourceDatabase** pro vyhledávání v názvech databází a **sourceOrganization** pro vyhledávání v názvech organizací. Tento příkaz je rychlejší alternativou příkazu **WDIcache(),** pokud uživatel chce najít informace o ukazateli s předem známým názvem (nebo částí názvu), nebo data z preferovaného zdroje.

K samotnému vyhledání dat pomocí knihovny WDI slouží následující příkaz:

```
> WDI (kód státu, kód ukazatele, …)
```

Seznam dostupných parametrů pro příkaz **WDI()** je vypsán v tabulce 5.

Tabulka 5: Seznam dostupných parametrů pro příkaz WDI()

Parametr	Požadován	Typ	Hodnoty	Popis
country	ANO	textový řetězec nebo vektor		Vektor kódů států. Pokud má hodnotu „all“, příkaz vrátí data pro všechny dostupné státy.
indicator	ANO	textový řetězec		Textový řetězec, který určuje požadovaný ukazatel.
start	NE	číslo		Určuje rok prvního záznamu dat.
end	NE	číslo		Určuje rok posledního záznamu dat.
extra	NE	logická hodnota	TRUE FALSE	Pokud má hodnotu TRUE, příkaz vrátí dodatečné proměnné (region, io3 kód). Primární hodnota FALSE
cache	NE	list	NULL	Určuje seznam ukazatelů. Pokud je hodnota NULL, je použit základní zdroj.

Zdroj: https://cran.r-project.org/web/packages/WDI/WDI.pdf

3.4.3 Ukázka práce s knihovnou WDI

V této podkapitole je ukázáno získání dat o hrubém domácím produktu na obyvatele, emisích oxidu uhličitého na obyvatele, nákladech spojené se založením podnikání na obyvatele a spotřeba alkoholu na obyvatele pomocí knihovny WDI. Dále jsou tyto státy rozděleny podle úrovně příjmů a poté vykresleny do boxplotu (nebo také krabicového grafu) pomocí příkazů programu R.

Nejdříve soubor příkazů pro získání požadovaných dat pro všechny státy.

```
> hdp <- WDI (country = "all", indicator = "NY.GDP.PCAP.KD", start = 2015, end = 2015, extra = TRUE)

> co2 <- WDI (country = "all", indicator = "EN.ATM.CO2E.PC", start = 2014, end = 2014, extra = TRUE)

> startup <- WDI (country = "all", indicator = "IC.REG.COST.PC.ZS", start = 2015, end = 2015, extra = TRUE)

> alcohol <- WDI (country = "all", indicator = "SH.ALC.PCAP.LI", start = 2015, end = 2015, extra = TRUE)
```

Pro argument **country** je zde použita hodnota **all**, díky které příkaz vrátí data pro všechny dostupné státy. Data o emisích oxidu uhličitého na obyvatele jsou jako jediné z roku 2014, neboť data z roku 2015 nebyla k 22. 5. 2018 dostupná. Argument **extra** má hodnotu TRUE, kvůli dostupnosti sloupce **income**, podle kterého jsou následně rozděleny státy do různých skupin dle úrovně jejich příjmů.

Struktura získaných dat je k vidění na obrázku 13 (jednotlivé proměnné se liší pouze sloupcem s hodnotami ukazatele, který je pojmenovaný jako kód daného ukazatele).

Obrázek 13: Struktura proměnné hdp

```
> str(hdp)
'data.frame':   264 obs. of  11 variables:
 $ iso2c         : chr  "1A" "1W" "4E" "7E" ...
 $ country       : chr  "Arab World" "World" "East Asia & Pacific (excluding hi$
 $ NY.GDP.PCAP.KD: num  6365 10287 5494 8945 1603 ...
 $ year          : num  2015 2015 2015 2015 2015 ...
 $ iso3c         : Factor w/ 304 levels "ABW","AFG","AFR",..: 8 297 81 84 239 6$
 $ region        : Factor w/ 8 levels "Aggregates","East Asia & Pacific",..: 1 $
 $ capital       : Factor w/ 212 levels "","Abu Dhabi",..: 1 1 1 1 1 10 2 80 16$
 $ longitude     : Factor w/ 212 levels "","-0.126236",..: 1 1 1 1 1 70 187 195$
 $ latitude      : Factor w/ 212 levels "","-0.229498",..: 1 1 1 1 1 151 100 12$
 $ income        : Factor w/ 5 levels "Aggregates","High income",..: 1 1 1 1 1 $
 $ lending       : Factor w/ 5 levels "Aggregates","Blend",..: 1 1 1 1 1 5 5 4 $
> 
```

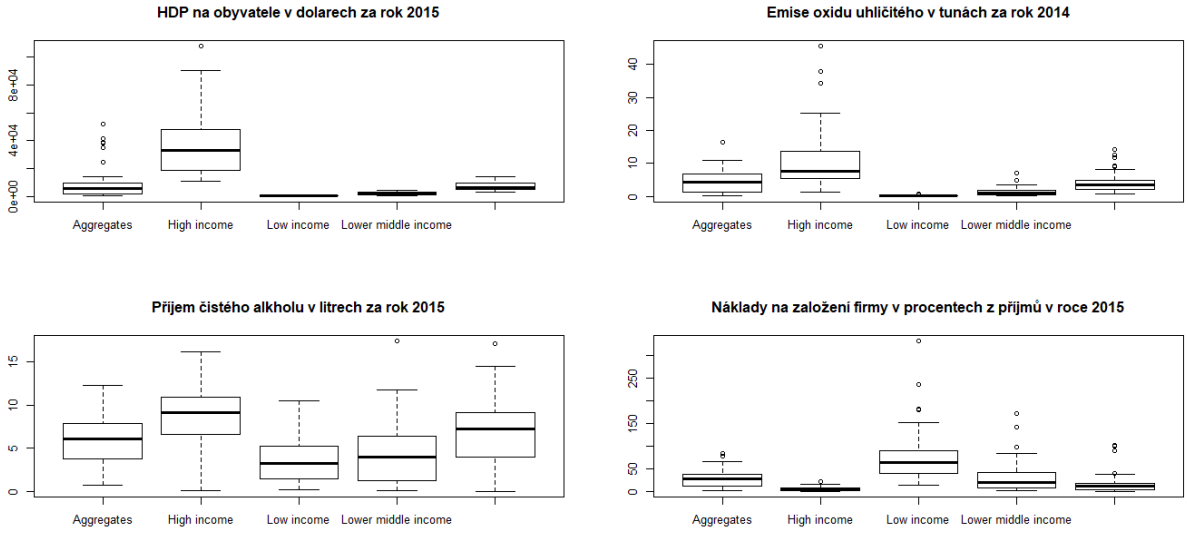
Zdroj: vlastní tvorba v programu R

Pro vykreslení boxplotů v programu R se používá následujících příkazů:

```
> par (mfrow = c (2,2))

> boxplot (hdp$NY.GDP.PCAP.KD~hdp$income, main = "HDP na obyvatele v dolarech za rok 2015")

> boxplot (co2$EN.ATM.CO2E.PC~co2$income, main="Emise oxidu uhličitého v tunách za rok 2014")

> boxplot (alcohol$SH.ALC.PCAP.LI~alcohol$income, main="Příjem čistého alkholu v litrech za rok 2015")

> boxplot (startup$IC.REG.COST.PC.ZS~startup$income, main="Náklady na založení firmy v procentech z příjmů v roce 2015")
```

Zde použitý příkaz **boxplot()** obsahuje dva argumenty. První argument je vzorec, který před tildou (~) obsahuje vektor hodnot, které chce uživatel vykreslit (v tomto případě hodnoty ukazatelů), a za tildou obsahuje vektor, který obsahuje hodnoty, podle kterých se hodnoty z prvního vektoru rozdělí do skupin (v tomto případě úroveň příjmů). Vektor rozdělující hodnoty do skupin je většinou typu **factor**, ale může se použít i vektor obsahující čísla či text. Druhý argument **main** vykresluje do grafu hlavní nadpis. Výsledné grafy je možné vidět na obrázku 14.

Obrázek 14: Krabicové grafy



Zdroj: vlastní tvorba v programu R

Nyní je ukázána jiná cesta pro získání dat o ukazatelích států rozdělený dle příjmů (samozřejmě státy mohou být stejným způsobem rozděleny například podle polohy nebo regionu). Nejdříve získání informací o státech pomocí příkazu **WDIcache()** a převedení proměnné na typ data frame.

```
> country <- as.data.frame (WDIcache()$country)
```

Následuje uložení států do nových proměnných v závislosti na hodnotě v sloupci **income**.

```
> lowinc <- country [country$income == "Low income",]
> lowmidinc <- country [country$income == "Lower middle income",]
> uppermidinc <- country [country$income == "Upper middle income",]
> highinc <- country [country$income == "High income",]
```

Získání dat pomocí funkce **WDI()** (ukázka pouze pro získání dat o hrubém domácím produktu):

```
> hdplow <- WDI (country = lowinc$iso2c, indicator = "NY.GDP.PCAP.KD", start = 2015, end = 2015, extra = TRUE)
> hdplowmid <- WDI (country = lowmidinc$iso2c, indicator = "NY.GDP.PCAP.KD", start = 2015, end = 2015, extra = TRUE)
```

```
> hdpuppermid <- WDI (country = uppermidinc$iso2c, indicator =
"NY.GDP.PCAP.KD", start = 2015, end = 2015, extra = TRUE)

> hdphigh <- WDI (country = highinc$iso2c, indicator = "NY.GDP.PCAP.KD",
start = 2015, end = 2015, extra = TRUE)
```

Příkaz pro vykreslení boxplotu by místo vzorce použitém v předchozím případě jako argument vykreslovaných dat obsahoval seznam vektorů obsahující tato data. Výsledný příkaz vypadá následovně:

```
> boxplot (x = c (hdplow$NY.GDP.PCAP.KD, hdplowmid$NY.GDP.PCAP.KD,
hdpuppermid$NY.GDP.PCAP.KD, hdphigh$NY.GDP.PCAP.KD), main = "HDP
na obyvatele v dolarech za rok 2015")
```

Pokud uživatel chce získat data pro všechny státy, a poté tato data rozdělit dle různých charakteristik států, je lepší použít předchozí postup. Pokud ovšem chce uživatel pracovat pouze s daty pro státy s určitou charakteristikou, druhý postup je rychlejší a zároveň úspornější, co se týče využití paměti a výkonu počítače.

4 Závěr

Na závěr této práce jsou zdroje ekonomických dat popsané v teoretické části a knihovny programu R popsané v praktické části porovnány z několika hledisek.

Z hlediska rozsahu dat, které zdroj může poskytnout uživateli, je na prvním místě Wolfram|Aplha, který je vytvořen spíše jako vyhledávač odpovědí na zadané dotazy uživatelů z více oblastí než jen ekonomické. Další v pořadí je Quandl, který sdružuje na jednom místě stovky databází, ze kterých čerpá data pro uživatele. Následují World Bank Data, IMF Data a United Nations Data, které jsou spravovány organizacemi aktivními po celém světě a poskytují data z mnoha oblastí lidské činnosti, která jsou dostatečná pro drtivou většinu uživatelů. OECD Data, Eurostat a Český statistický úřad jsou zdroje omezené na určité geografické či politické skupiny států. WTO Data je pro změnu zdroj poskytující pouze data o mezinárodním obchodu. Nakonec The World Factbook, který poskytuje spíše shrnutí informací o jednotlivých státech, což je ovšem prezentace informací, které někteří uživatelé mohou vyhledávat.

Dalším hlediskem je množství finančních prostředků nutných pro získání dat. Většina zde popsaných zdrojů poskytuje data pro nekomerční účely zdarma. Výjimkou jsou Wolfram|Aplha, u kterého je nutné pro jakékoliv stažení dat nebo používání služby API zaplatit poplatek, a Quandl, kde je přístup k některým databázím omezen nutností poplatku, ale většina databází je přístupná zdarma a pro většinu uživatelů jsou data dostupná z těchto databází dostačující.

Z hlediska zpracování webových stránek a obtížnosti získávání požadovaných dat je samozřejmě pro uživatele ovládající český jazyk a hledající informace o České republice nejlepší volbou databáze Českého statistického úřadu, kde nevzniká jazyková bariéra pro anglicky nemluvící uživatele. Dále jsou velice krásně zpracované, přehledné a lehké k orientaci internetové stránky Světové banky a The World Factbook. Na opačném spektru jsou z tohoto hlediska stránky IMF a Quandlu.

Ze zdrojů, které poskytují službu API, je na velice dobré úrovni v používání a zpracování dokumentace pro tuto službu od Světové banky a Quandlu. Quandl má navíc výhodu ve sjednocení struktury dotazů pro stovky zdrojových databází od různých poskytovatelů.

Z knihoven popsaných v praktické části jsou v určitých maličkostech lepší knihovny WDI a wbstats oproti zbylým dvěma. Obě zpřístupňují službu API Světové

banky, a proto poskytují velice podobné možnosti a funkce, tudíž je těžké mezi těmito dvěma knihovnami rozhodnout, která je lépe zpracovaná. Knihovna eurostat je na podobné úrovni z hlediska funkcionality jako dvě dříve zmíněné knihovny, poskytuje ovšem pouze data pro omezené množství států. Naopak knihovna Quandl poskytuje zdaleka největší množství dat, ale z hlediska praktičnosti nedosahuje kvalit ostatních knihoven (například není možné získat data o jednom ukazateli pro více států v jednom dotazu, ale je nutné zadat dotaz pro každý stát zvlášť).

I. Summary a keywords

The main target of this bachelor thesis is to describe acquiring economic data from the online sources and show examples of processing the data in program R.

The theoretical part of this thesis presents ten online sources of economic data and basic information about program R and programming language R. The sources that are described here are The World Bank Data, United Nations Data, IMF Data, WTO Data, OECD Data, The World Factbook (by CIA), Wolfram|Aplha, Eurostat, Czech Statistical Office and Quandl. The operator of each source is briefly introduced and then the possibilities of acquiring data for each source are described.

The practical part of the thesis presents four libraries in program R which are focused on acquiring data. The libraries which are described here are WDI, eurostat, wbstats and Quandl. The possibilities of acquiring data and practical usage of methods the library provides in program R are shown for each library. Then the examples of transforming and processing the acquired data in program R are described.

Keywords: sources of economic data, acquiring data, WDI, eurostat, wbstats, Quandl, program R

II. Seznam použitých zdrojů

Český statistický úřad. (22. Květen 2018). *O ČSÚ | ČSÚ*. Načteno z Český statistický úřad | ČSÚ: https://www.czso.cz/csu/czso/o-csu

Eurostat. (22. Květen 2018). *About this service - Eurostat*. Načteno z Home - Eurostat: http://ec.europa.eu/eurostat/web/json-and-unicode-web-services

Chambers, J. M. (2008). *Software for Data Analysis: Programming with R.* New York: Springer.

Mezinárodní měnový fond. (22. Květen 2018). *About the IMF*. Načteno z IMF -- International Monetary Fund Home Page: https://www.imf.org/en/About

Organizace pro ekonomickou spolupráci a rozvoj. (22. Květen 2018). *About the OECD - OECD*. Načteno z OECD.org - OECD: http://www.oecd.org/about

Organizace pro ekonomickou spolupráci a rozvoj. (22. Květen 2018). *API Documentation*. Načteno z OECD Data: https://data.oecd.org/api/sdmx-json-documentation/

Organizace spojených národů. (2018, Květen 22). *UNdata | about us*. Retrieved Květen 22, 2018, from UNdata: http://data.un.org/Host.aspx?Content=About

Světová banka. (2018, 5 22). *About the World Bank*. Retrieved from World Bank Group - International Development, Poverty, & Sustainability: http://www.worldbank.org/en/about

Světová banka. (2018, 5 22). *API: Basic Call Structure – World Bank Data Help Desk*. Retrieved from World Bank Group - International Development, Poverty, & Sustainability: https://datahelpdesk.worldbank.org/knowledgebase/articles/898581-api-basic-call-structure

Světová obchodní organizace. (22. Květen 2018). *WTO | About the organization*. Načteno z World Trade Organization - Home page: https://www.wto.org/english/thewto_e/thewto_e.htm

Ústřední zpravodajská služba USA. (22. Květen 2018). *About CIA — Central Intelligence Agency*. Načteno z Welcome to the CIA Web Site — Central Intelligence Agency: https://www.cia.gov/about-cia

Wikimedia Foundation. (22. Květen 2018). *Eurostat – Wikipedie*. Načteno z Wikipedie, otevřená encyklopedie: https://cs.wikipedia.org/wiki/Eurostat

Wikimedia Foundation, Inc. (22. Květen 2018). *Wolfram Alpha - Wikipedia*. Načteno z Wikipedia, the free encyclopedia: https://en.wikipedia.org/wiki/Wolfram_Alpha

Wikimedia Foundation, Inc. (22. Květen 2018). *API – Wikipedie*. Načteno z Wikipedie, otevřená encyklopedie: https://cs.wikipedia.org/wiki/API

Wikimedia Foundation, Inc. (22. Květen 2018). *Quandl - Wikipedia*. Načteno z Wikipedia, the free encyclopedia: https://en.wikipedia.org/wiki/Quandl

Wolfram Alpha LLC. (22. Květen 2018). *Wolfram|Alpha Simple API: Reference & Documentation*. Načteno z Wolfram|Alpha: Computational Intelligence: https://products.wolframalpha.com/simple-api/documentation/

III. Seznam obrázků

IV. Seznam tabulek