Sem vložte zadání Vaší práce.

FAKULTA
INFORMAČNÍCH
TECHNOLOGIÍ
ČVUT V PRAZE

Diplomová práce

Hledání a analýza proměnných souvisejících s TFR za využití metodiky BI

Bc. Jaroslav Jasenovský

Katedra softwarového inženýrství
Vedoucí práce: PhDr. Ing. Tomáš Evan, Ph.D.

2. ledna 2023

Poděkování

Chtěl bych poděkovat PhDr. Ing. Tomáši Evanovi, Ph.D. za vedení mé diplomové práce, cenné rady a odborný dohled. Děkuji také své rodině a přítelkyni za podporu při tvorbě práce a při hledání nových faktorů ovlivňujících TFR. Nakonec bych rád poděkoval Mgr. Janče Menšíkové za finální kontrolu.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principu při přípravě vysokoškolských závěrečných prací.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona, ve znění pozdějších předpisu. V souladu s ust. § 2373 odst. 2 zákona č. 89/2012 Sb., občanský zákoník, ve znění pozdějších předpisu, tímto uděluji nevýhradní oprávnění (licenci) k užití této mojí práce, a to včetně všech počítačových programu, jež jsou její součástí či přílohou a veškeré jejich dokumentace (dále souhrnně jen „Dílo“), a to všem osobám, které si přejí Dílo užít. Tyto osoby jsou oprávněny Dílo užít jakýmkoli zpusobem, který nesnižuje hodnotu Díla a za jakýmkoli účelem (včetně užití k výdělečným účelum). Toto oprávnění je časově, teritoriálně i množstevně neomezené. Každá osoba, která využije výše uvedenou licenci, se však zavazuje udělit ke každému dílu, které vznikne (byť jen zčásti) na základě Díla, úpravou Díla, spojením Díla s jiným dílem, zařazením Díla do díla souborného či zpracováním Díla (včetně překladu) licenci alespoň ve výše uvedeném rozsahu a zároveň zpřístupnit zdrojový kód takového díla alespoň srovnatelným zpusobem a ve srovnatelném rozsahu, jako je zpřístupněn zdrojový kód Díla.

V Praze dne 2. ledna 2023 . . . . . .

Odkaz na tuto práci

Jasenovský, Jaroslav. *Hledání a analýza proměnných souvisejících s TFR za využití metodiky BI*. Diplomová práce. Praha: České vysoké učení technické v Praze, Fakulta informačních technologií, 2023.

Abstrakt

Cílem diplomové práce bylo nalezení silných faktorů ovlivňujících úhrnnou plodnost, a to za pomocí metodiky BI. Za tímto účelem byl navržen a vytvořen datový sklad, v němž se uchovávají zkoumané datové sady ve všech hlavních stavech, které během procesu zpracování pomocí technologií ETL nastanou.

Tato data byla následně pomocí pro tyto účely vytvořeného dashboardu a následně také komplexně pomocí vytvořeného skriptu analyzovány a vyhodnoceny z pohledu jejich vlivu na sledovanou veličinu. Pro větší dostupnost těchto dat širokému okolí byl zároveň daný analytický dashboard publikován na mnou vytvořené webové stránky.

Klíčová slova Úhrnná plodnost, lineární regrese, datové sklady, Byznys inteligence, ELT

Abstract

The aim of the thesis was to find the strong factors influencing the cumulative fertility using the BI methodology. For this purpose, a data warehouse was designed and created in which the investigated datasets are stored in all the main states that occur during the processing using ETL technologies.

These data were then analyzed and evaluated in terms of their impact on the variable under study using a dashboard created for this purpose and then comprehensively using a script created. In order to make this data more accessible to the general public, the analytical dashboard was also published on a website created by me.

Keywords Total Fertility Rate, linear regression, data warehouse, business intelligence, ETL

Obsah

Seznam obrázků

Seznam tabulek

Úvod

Problematika klesající porodnosti je problémem ve všech vyspělých zemích již desítky let. A díky tomu, že v mnoha dalších zemích dochází k přerodu, je tato situace již globální, a pokud se nesrovná, bude populace stále více stárnout, až dojde do stavu, kdy dnešní sociální systém, jak ho nyní známe, již nebude udržitelný a celková populace začne klesat.

Díky tomu, že tuto problematiku zkoumáme a hledáme, jaké vlivy na tuto hodnotu působí, ať kladně či záporně, můžeme tak nalézt zajímavé vztahy, díky kterým by šla tato situace zvrátit a dostat hodnotu TFR zpátky nad 2,1 bodu. To pomůže jak našim starším generacím, pro které bude existovat stále dostatečně velká aktivní skupina v populaci, aby bylo možné pomáhat lidem v důchodovém věku, ale hlavně je to i pozitivní vyhlídka pro nás mladší generace, že bude stále šance se nějakého důchodu dožít. To souvisí i s mojí motivací pracovat na tomto tématu. Chci pomoci zkoumat tuto problematiku a rozšířit povědomí o ní, aby se naše populace stala udržitelnou, a zajistil jsem tak budoucnost nejen nám mladším ročníkům, ale také generacím budoucím, do které budou patřit i mé děti.

Pro řešení tohoto problému jsem se rozhodl využít datový sklad, který je určen k uložení velkého množství historizovaných dat a k vytváření následných analýz. Díky tomu, že se v datových skladech využívají dimenzionální databáze, lze tak snadno zkoumat porodnost jak v různých částech světa, v různých časových úsecích, ale také s různými veličinami, které chceme s úhrnnou plodností porovnávat. Díky tomu lze tuto práci využít i v budoucnu pro další analýzy různých dalších faktorů, které stačí pouze přidat do datového skladu a buď vytvořit samotný nový datamart a vizuál dle našich preferencí, či přidat tuto novou položku do dimenze faktorů a následně ji srovnávat s ostatními v již připraveném dashboardu.

Tento dashboard je vhodný pro přehledné zobrazení výsledků dané analýzy a umožňuje snadné procházení skrze různé dimenze, a díky tomu dochází k porovnání dané veličiny například v prostoru a čase. Dashboard bude pro-

pojen s webovou stránkou, kde budou výsledky této práce zveřejněny, a bude tak umožněno ostatním uživatelům s těmito výsledky přehledně pracovat a zkoumat vzájemné vztahy.

Kapitola 1

Cíl práce

Cílem této práce je provést čtenáře problematikou TFR a pojmy s ní související, dále také seznámit ho s dopady na lidstvo a možný budoucí vývoj populace. S tím budou také představeny silné korelanty, které tuto veličinu ovlivňují a jsou statisticky ověřeny.

Dále bude představena statistická metodologie, jež je v této práci použita, a která nám poskytuje statisticky průkazné důkazy a vzájemné závislosti zkoumaných hodnot na TFR a jejich silných korelantech.

Čtenář bude také proveden principy Business intelligence, které jsou využity pro datové zpracování a následné zkoumání statistiky. Také jsou využity pro přehledné zobrazení výsledků.

V praktické části je cílem navržení, sestavení a naplnění datového skladu, dále hledání vhodných datových sad s vybranými proměnnými, které mohou mít vliv na sledovanou veličinu. K účelu sběru dat jsou využity veřejné databáze, které taková data poskytují, a v případě, že pro danou zkoumanou proměnnou nejsou vhodná data nalezena, je využito Google Trends, kde je využita jejich četnost vyhledávání, což má samo o sobě vypovídající hodnotu, zejména ve vyspělém světě, kde je internet součástí každodenního života.

Tyto veličiny jsou následně vloženy do datového skladu a pomocí výše zmíněnné technologie jsou zkoumány vzájemné vlivy na TFR. U všech proměnných, kde se ukáže, že je zde nějaký vzájemný vliv, se následně tato teorie ověřuje podrobnější statistickou metodou, na základě které se tato hypotéza potvrdí, či vyvrátí.

Přínosem práce je vybudování prostředí pro přehledné testování veličin souvisejících s TFR, které mohou být následně využity k možné podpoře růstu úhrnné plodnostia a také k nalezení zajímavých vztahů mezi určitými korelanty a TFR. Toto vše bude ve finále prezentováno v interaktivní podobě na webových stránkách, díky čemuž bude možné tyto vztahy prezentovat širokému spektru zájemců o tuto problematiku.

Kapitola 2

Fertility rate

Fertility rate, česky míra plodnosti, je jeden z hlavních faktorů, který je zkoumán sociologicko-ekonomickou společností. Je to hlavně díky tomu, že má velký dopad na budoucí vývoj sociologických systémů a ekonomiky v dané zemi.

Vlivem klesající plodnosti, zejména ve vyspělých zemích, ale i v globálním měřítku, celková společnost stárne a to značně mění poměr ekonomicky aktivních lidí vůči těm, kteří jsou již v důchodovém věku. To ovšem v zemích, kde je sociální systém a jsou zde vypláceny důchody, značně zvyšuje částku potřebnou na jejich vyplácení. Tu stát vybírá na sociálních daních od aktuálně pracovně aktivní skupiny lidí, jejíž poměr se vůči té neaktivní, díky klesající fertility rate, neustále snižuje.

To má za následek, že je na důchody investováno stále větší procento z částky vybrané z daní a díky tomu se také zmenšuje část peněz, která může být investována například do dopravy, zdravotnictví nebo školství.

2.1 Total Fertility Rate

Celková míra porodnosti nám udává, kolik v průměru připadá dětí na každou ženu na konci jejího produktivního věku, což je v dnešní době statisticky mezi 15 a 49 lety.

Hraniční hodnotou pro zachování populace je považována hodnota 2,1, kdy na každou ženu vychází více jak 2 děti. Ovšem u méně vyspělých států, převážně na africkém kontinentu, je tato hodnota výrazně vyšší. Touto fází vývoje si prošly všechny země a s tím, jak se vyvíjí, se v dané zemi mění i tato veličina.

Pokud je TFR dlouhodobě pod touto hranicí, zejména pokud je pod hodnotou 1,7, tak zde populace vymírá a může se dostat do stavu, ze kterého již není, bez vnějších vlivů jako je například imigrace, návratu. Pro výpočet úhrnné plodnosti se používá vzorec:

$$TFR = \sum_{k=15}^{49} \frac{N_k^v}{P_k^z},$$

kde N_k^v představuje živě narozené děti pro každou věkovou skupinu žen ve věku k (15–49 let) a P_k^z představuje populace jednotlivých věkových skupin žen ve věku k ke střednímu stavu obyvatelstva.

Tento výpočet vyžaduje velmi specifická a přesná data o počtu narozených dětí v jednotlivých věkových skupinách. Z toho důvodů je často velmi náročné až nemožné vypočítat hodnoty TFR ve vzdálenější minulosti. Naštěstí díky zvyšujícímu se povědomí o této problematice a nárůstu průzkumů, které se této věci věnují, se kvalita a sběr dat ve většině zemí světa zlepšuje, díky čemuž jsme schopni získávat přesnější hodnoty a docházet tak ke korektnějším výsledkům, z nichž pak následně vyvozujeme naše závěry. [2]

2.2 Významné korelanty ovlivňující TFR

Úhrnná plodnost je veličina, která je ovlivněna velkým množstvím faktorů, a i já se snažím v této práci prozkoumat nové či méně známé, který tuto hodnotu ovlivňují. Ovšem ty nejvýznamnější činitele jsou již velice dobře prozkoumány, a tak považuji za správné je v této práci také zmínit, neboť jsem se při hledání nových proměnných často inspiroval i těmito známými faktory. Zároveň jsou mnohé tyto proměnné také součástí množiny hodnot, kterou pomocí datového skladu zkoumám, abych na nich mohl ověřit, zda použitá analýza dává očekávané hodnoty, a ověřil tak její funkčnost.

Tímto tématem se zabývala také práce *Hledání a práce s veličinami souvisejícími s TFR* [10], kterou velmi pěkně zpracoval kolega Daniel Brotz. Ve své webové aplikaci prozkoumal velké množství proměnných a odvedl tím tak velký kus práce. Já jsem k tomuto tématu přistoupil trochu jinak a více než na kvantitu jsem se soustředil na vlastní hloubku, kdy jsem využil právě metodik BI k tomu, abych mohl zkoumat zajímavé detaily v určitých zemích a letech, ale zároveň jsem využíval kompletní dostupné datové sady, abych měl, co nejširší spektrum zemí a let, ve kterých lze tato data porovnávat. Také jsem zvolil publikaci na webové stránky, díky čemuž jsem umožnil zasáhnout větší cílovou skupinu, která má tak možnost toto téma zkoumat. V poslední řádě jsem díky své zvolené metodice zvolil kompletně jiný přístup k návrhu a zpracování dat.

Mezi tyto nejvýznamnější faktory patří vzdělání (zejména u žen), antikoncepce, ženská zaměstnanost, penze a celkové sociální výdaje. V díle *Human fertility in relation to education, economy, religion, contraception, and family planning programs*[3] je většina těchto faktorů zkoumána a věsměs potvrzuje jejich vliv na TFR. Také ovšem poukazuje na výrazný rozdíl ve vlivu mezi jednotlivými zkoumanými regiony (západní Evropa, východní Evropa, Latinská

Amerika, Arábie, Asie, Subsaharská Afrika), kdy v některých regionech má daný faktor velmi velký vliv a v jiné může být malý až žádný vliv, a dokonce může mít úplně opačný vliv, tedy že místo aby TFR například zvyšoval, tak ho snižuje a naopak.

Toto může do jisté míry souviset s vyspělostí daných zemí v regionu, neboť tyto rozdíly jsou nejčastěji vidět mezi subsaharskými státy a Evropou, ale může na to mít vliv i kultura a další sociologické aspekty. S kulturou jde ruku v ruce i náboženství, které je v *Human fertility in relation to education, economy, religion, contraception, and family planning programs*[3] sledován jako samostatný ovlivňující faktor, a i když jsou ze získaných dat vidět jisté vlivy, tak statistické hodnoty nám pro tento faktor nedávají dost silný důkaz, abych ho mohli potvrdit, a tak je třeba toto téma ještě lépe prozkoumat.

Kromě rozdílného vlivu významných hodnot v různých regionech je také prokázaný vliv v období, kdy na sebe dané veličiny působí. V tomto případě je vidět rozdílné výsledky pro testované období před rokem 2000 a po tomto roce, což je popsáno a ověřeno v díle *Influence of women's workforce participation and pensions on total fertility rate: a theoretical and econometric study* [8]. Autoři věří, že to může souviset se změnou příležitostí, které najednou byly zejména ve východní Evropě v 90. letech, kdy najednou měli obyvatelé více možností cestovat a studovat, a tak se první dítě často odkládalo.

Také je v této práci dobře znázorněno, že jednotlivé veličiny často nejsou hlavními hybately, ale že se jedná o komplexní problém, a že při kombinaci několika různých faktorů najednou se jejich celkový vliv na TFR může významně změnit.

2.2.1 Antikoncepce

Zcela zřejmý činitel je antikoncepce, kdy dochází k vědomé ochraně před početím během pohlavního styku. Tato ochrana je spojena s ekonomickou vyspělostí dané země, kde ve více vyspělých zemích je antikoncepce dostupnější a je zde i širší spektrum možností. Ovšem jak je uvedeno v průzkumu *Does age-adjusted measurement of contraceptive use better explain the relationship between fertility and contraception?* [4], tak samotný vliv antikoncepce postupně klesá, i když se stále jedná o významný ovlivňující faktor, není jeho vliv tak silný, jako býval před 20 lety. A díky tomu, že se průzkum zaměřil také na rozdíl v různých věkových skupinách, je z něho vidět i změna v užívání v různých skupinách za dobů testovaného období.

Síla tohoto faktoru je potvrzena i v *Human fertility in relation to education, economy, religion, contraception, and family planning programs*[3], kde také poukazují na to, že hlavně v Evropě je jeho vliv nižší. To může být také tím, že obzvláště hormonální antikoncepce, která je celkově velmi populární ve vyspělých zemích, je čím dál častěji očerňována za svůj vliv na zvyšující se neplodnost. A i když je pravda, že po dlouhodobém užívání má jistý post efekt, kdy se může případné početí pozdržet, tento nebývá nijak výrazný, a jak

uvádí studie *Return of fertility after discontinuation of contraception* [5], neexistuje vědecky ověřená spojitost mezi hormonální antikoncepcí a následnou neplodností.

2.2.2 Vzdělání a zaměstnání

Míra využití této ochrany je ovlivněna také vzděláním, kdy lidé s vyšší mírou vzdělání mají větší tendenci tyto prostředky využívat, protože díky tomu o ni získávají větší povědomí a jsou si také více vědomi toho, jak by dítě ovlivnilo jejich aktuální život. Také samotná doba vzdělání výrazně ovlivňuje plodnost, neboť ženy oddalují první dítě až po studiu, a tak přichází o svá nejplodnější léta. Studium na vysoké škole, tak samo o sobě dokáže vzít klidně deset a více nejplodnějších let u žen, které se ji rozhodly studovat (samozřejmě v případě, že neotěhotní již během studia).

Vzdělání dále ženám otevírá lepší pracovní příležitosti, což má za následek další oddalování prvního dítěte. Ve vyspělých zemích bývá často kladen důraz na kariéru a to nejde příliš ruku v ruce s mateřskou dovolenou a celkovou péči o dítě. I tento faktor má často za následek oddalování prvního početí, což snižuje možný počet dětí, které může daná žena mít, a zároveň snižuje šance na to mít vůbec nějaké.

Tyto faktory se dnes dostávají do popředí i v zemích Arabského světa, kde dříve z kulturních a náboženských důvodů byly tyto možnosti ženám téměř nedostupné, což je jeden ze zmíněných faktorů v díle *Contributing factors to the total fertility rate declining trend in the Middle East and North Africa* [7]. Je tedy vidět, že postupně dochází v celém světě ke kulturnímu přerodu a vývoji zemí, který má za následek postupný pokles porodnosti.

2.2.3 Sociální dávky

Spolu s penzijním zajištěním od státu jsou tyto faktory dobře popsány v publikaci *Influence of women's workforce participation and pensions on total fertility rate: a theoretical and econometric study* [8], kde je popsáno, jak v historii fungovaly děti jako způsob zajištění na důchod a jak se se tato potřeba v zemích s nastaveným státním penzijním systémem snižuje. A to i přes státní podporu na dítě a další mateřské benefity, které mohou matky čerpat, neboť mohou jen stěží konkurovat příležitostem, kterým se ženám naskýtají v zemích vyspělého světa. Prorodinné a další dávky mají velký význam ve zvyšování TFR, ale tento vliv je velmi citlivý a může být snadno převážen jinými sociálními dávkami, zejména pak penzijním zajištěním, které s klesající hodnotou TFR a stárnoucí populací bude zcela neudržitelné.

2.2.4 Urbanizace

Přesun obyvatel do měst je často spojen s výraznými ekonomickými a sociálními změnami, a to zejména díky lepší dostupnosti všech aspektů, které

jsou hlavními ovlivňovateli těchto parametrů. Mezi ně lze započítat například zdravotní péči, snazší přístup k lepšímu vzdělání, ale i mnohem lepší pracovní příležitosti. A jak je již z tohoto vybraného vzorku vidět, tyto věci často samy o sobě mají silný vliv na TFR, a díky tomu lze vyvodit, že i samotná urbanizace má pak na změnu v porodnosti svůj podíl. Jak vyplývá z *Urbanization and Fertility Decline: Cashing in on Structural Change* [9], tak samotná urbanizace má značný efekt na hodnotu TFR, kde výzkum udává průměrný rozdíl až 1,5 dítěte na ženu. Ovšem také z něho plyne, že tento vliv je spíše nepřímého charakteru, což značně koresponduje s tím, co jsem psal v úvodu této podkapitoly. Sama urbanizace není činitel, ale dává mnohem větší příležitost vyniknout dalším hlavním vlivům.

Ovšem tento vliv také závisí na samotné vyspělosti dané země, neboť obzvláště dříve byl velký rozdíl v tom, jaký vliv měl tento faktor na porodnost mezi zeměmi afrického kontinentu a třeba zeměmi v Jižní Americe či samotné Evropě, což je lépe popsáno v díle *Rural-Urban Differences in Fertility: An International Comparison* [11]. V některých afrických zemích dokonce byla hodnota porodnosti v městech větší, což ovšem úplně nepopírá to, že hlavní vliv urbanizace je v tom, že podporuje ostatní vlivy, které ovšem v afrických zemích v 50. a 60. letech nebyly ani ve městech příliš rozvinuty, a tedy jejich vliv nemohl být samotnou urbanizací podpořen.

2.3 Dělení TFR podle fáze vývoje

Tyto fáze jsou závislé na socioekonomickém přerodu dané země a postupně jimi prochází všechny země světa. Prvními byly zejména země na území Evropy, kde tento přerod byl ovlivněn hlavně průmyslovou revolucí a následným rychlým ekonomickým růstem.[6]

Země, kde hodnoty TFR nabývají hodnot 5 a více, se nachází v první fázi, kdy ještě nebyla započata revoluce v plodnosti. Tyto země se aktuálně nejvíce soustředí na území Afriky a jsou charakteristické nízkou mírou vzdělání, vysokou úmrtností a malou sociální péčí. [1] V těchto zemích ovšem i tak dochází k pozvolnému poklesu TFR a přechodu do další fáze. Nejvyšší hodnotu má Nigérie, kde úhrnná plodnost dosahuje přibližně hodnoty 6,7.

V druhé fázi jsou země s hodnotami TFR mezi 2,1 a 5. Zde ještě populace stále narůstá, často vlivem snižující se úmrtnosti může narůstat razantně. Jsou to země, kde dochází k postupnému přerodu v klíčových parametrech ovlivňující plodnost, zejména pak zvyšující se možnosti vzdělání a následného pracovního uplatnění má velký vliv na oddalování prvního dítěte. V těchto zemích celkový přerod pokračuje a v závislosti na různých faktorech (politika, kultura, ekonomika) pak rychle či pozvolně směřují k poslední fázi. V této fázi se nachází například Indie, sever Afriky, Argentina či státy Arabského poloostrova. [12]

V poslední fázi se nachází většina vyspělých států světa a jejich hodnota

TFR klesla pod mezní hodnotu 2,1; zde se buď ustálí v okolí nějaké hodnoty v závislosti na zvolených opatřeních v dané zemi, a nebo dále pozvolna klesá. To má za následek celkové stárnutí populace a následné vymírání. Nejhůře je na tom aktuálně Jižní Korea, která se nyní nachází na hodnotě přibližně 0,8.

Kapitola 3

Matematické metody

Data, která jsou využita v této práci, mají podobu časových řad, což jsou posloupnosti chronologicky uspořádaných hodnot, většinou s určitou frekvencí (měsíc, rok, …). Tyto řady lze pro účely analýzy reprezentovat jako náhodný vektor, který má následující definici.

Pro $n \in N$ uvažujme náhodné veličiny $X_1, ..., X_n$ na stejném pravděpodobnostním prostoru (Ω, F, P). Vektor $X = (X_1, ..., X_n)^T$ potom nazýváme náhodným vektorem na (Ω, F, P).

Tyto řady jsou často zatíženy nejistotou a jsou tedy nedeterministické, jejich chování tudíž nelze jednoznačně popsat matematickým vzorcem. [13]

3.1 Hypotéza

Máme-li tvrzení od rozdělení náhodného vektoru, jehož platnost není známá, pak takové tvrzení nazveme hypotézou. Mechanizmus, jak tuto platnost ověřit na základě pozorovaných hodnot X, se nazývá **testování hypotéz**.

Při testování se pracuje se dvěma hypotézami. Nulovou H_0 označujeme tvrzení, o kterém chceme rozhodovat. A druhou je opačné tvrzení vůči H_0, které nazýváme alternativní hypotézou H_A. Test nulové hypotézy H_0 proti alternativní hypotéze H_A je rozhodovací proces založený na hodnotě X, na jehož základě zamítáme nebo nezamítáme hypotézu H_0.

Při testování může dojít chybám dvojího druhu. **Chyba prvního druhu**, kdy zamítáme H_0, ačkoliv platí, a **chyba druhého druhu**, kdy H_0 nezamítáme, ačkoliv neplatí. Během testování nelze kontrolovat oba druhy chyb, a tak se testy sestavují tak, aby se minimalizovala chyba prvního druhu. H_0 potom volíme takové, že chyba prvního druhu je závažnější než chyba druhého druhu. Pravděpodobnost, že nastane chyba prvního druhu, je pak nejvýše rovna hladině významnosti testu α, kterou si sami zvolíme. Nejčastěji se pro α volí hodnoty 1 % či 5 %. Pravděpodobnost, že nastane druhá chyba, je neznámá.

Zamítnutí H_0 ve prospech H_A je tedy silný výsledek; jestliže zamítáme H_0, můžeme s velkou spolehlivostí tvrdit, že platí alternativa H_A, a říkáme, že tvrzení H_A je statisticky významné. Pokud H_0 nezamítáme, nazýváme tvrzení H_A statisticky nevýznamné. Hypotézu, kterou potřebujeme dokázat, tedy volíme jako alternativní hypotézu H_A. Je-li pak výsledkem testu zamítnutí H_0, víme, že H_A platí s pravděpodobností alespoň $1 - \alpha$.[14]

3.2 Testování hypotéz

Ve statistice se používá mnoho typů testů pro ověření hypotéz a pro různé případy. Základními reprezentanty, kteří se nejběžněji používají, jsou t-test a F-test. První zmíněný se využívá při zkoumání velikosti naměřených hodnot a porovnávají se v tom případě jejich průměry. Druhý test se nejčastěji používá v případě, že nás zajímá variabilita naměřených hodnot, a porovnáváme v tomto případě jejich rozptyly.

Dalším důležitým faktorem testování je i jaký typ testu použít. V tomto případě je třeba rozhodnout, jestli nás zajímá uzavřenost řešeného problému z obou stran, a použijeme tedy oboustrannou verzi testu, a nebo nám stačí pouze jednostranná varianta. Toto primárně závisí na tom, jak je sestavena hypotéza a jestli je tedy použita rovnost, nebo jestli nás zajímá, zda je jedna veličina větší či menší.

Poslední důležitým faktorem je také, jestli zkoumáme pouze jeden výběr, který následně porovnáváme s nějakou referenční hodnotou, nebo máme dva výběry, kdy se následně musíme rozhodnout, zda jsou tato data párová či nepárová.

Pro jednodušší orientaci při výběru testu u základnách případů poslouží následující seznam: [25]

- **Jednovýběrový** - jednovýběrový t-test (porovnáváme s referenčním průměrem)
- **Dvouvýběrový**
  - **Testování rozptylů** - dvouvýběrový F-test
  - **Nepárová data** - potřeba rozhodnout, zda jsou rozptyly shodné, či odlišné, například F-testem
    - **Shodné** - dvouvýb. t-test pro nepárová data, se shodnými rozptyly
    - **Neshodné** - dvouvýb. t-test pro nepárová data, s odlišnými rozptyly

Při provádění těchto testů získáme několik hodnot. Těmi nejdůležitějšími jsou kritická hodnota, která nám udává bod zlomu, v němž se pro daný případ láme možnost zamítnutí a nezamítnutí nulové hypotézy.

Druhou hodnotou je p-hodnota, která má hodnoty v množině (0; 1) a dá se definovat jako pravděpodobnost, že nastane chyba prvního druhu, tedy že zamítáme chybně H_0. P-hodnotu tedy porovnáváme s hodnotou α, kterou jsme si zvolili. Je tedy na nás, jak nízkou hladinu pro zamítnutí H_0 zvolíme, ale jak jsem již zmínil výše, nejčastěji se pracuje s hodnotami 0,05 (5 %) a 0,01 (1 %). Platí tedy vztah:

- Pokud je p > α, pak H_0 nezamítáme.
- Jinak když je p < α, pak H_0 zamítáme ve prospěch H_A.

3.3 Linerární regrese

Lineární regrese se používá k získání předpisu příčinné souvislosti mezi náhodnými veličinami. Model lineární regrese má tvar

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

kde Y_i je závislá (vysvětlovaná) proměnná a X_i je nezávislá (vysvětlující) proměnná. Parametry (koeficienty) β_0 a β_1 jsou neznámé konstanty, musíme je tedy odhadnout z dat. β_0 je konstantní člen (udává posunutí přímky po ose y) a β_1 směrnice přímky (určuje sklon přímky), ϵ_i je reziduální (chybový) člen. [16]

Pro odhadnutí koeficientů tak, aby co nejlépe popisovaly naše data, využíváme **metodu nejmenších čtverců**, kde se minimalizuje součet čtverců všech reziduí. Díky tomu získáme odhad rozptylu chybového členu, což je důležité pro testování významnosti regresních koeficientů a určení konfidenčních intervalů.

Přímka $y = \beta_0 + \beta_1 x$ je přímka proložená metodou nejmenších čtverců souborem bodů $[x_1, y_1], [x_2, y_2], ..., [x_n, y_n]$, jestliže pro koeficienty a, b platí

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2,$$

kde $\widehat{y}_i$ je výsledek rovnice $\widehat{y}_i = \beta_0 + \beta_1 x_i$ po dosazení odhadů parametrů β_0 a β_1.

S takto sestavenými a získanými hodnotami můžeme následně vytvořit nulovou hypotézu ve tvaru:

$$H_0 : \beta_1 = 0$$

proti alternativní hypotéze

$$H_A : \beta_1 \neq 0.$$

V tomto případě má konfidenční interval podobu:

$$\beta_1 + t_{\alpha/2,n-2} \frac{\sqrt{s^2}}{\sqrt{\sum (x - \hat{x})^2}}.$$

V této rovnici je $t_{\alpha/2,n-2}$ hodnota Studentova t-rozdělení a s^2 je reziduální rozptyl, který získáme výpočtem následující rovnice

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y-\hat{y})^2.$$

Následně můžeme na základě p-hodnoty testu rozhodnout o platnosti nulové hypotézy a při jejím zamítnutím přijmout její alternativu, tedy že testovaná proměnná je dobrým prediktorem k druhé proměnné. [16]

3.4 Korelace

Korelační koeficient je další možnost, jak vyšetřit lineární míru závislosti mezi veličinami. Nejdříve je ovšem třeba definovat další pojmy, z kterých samotná korelace vychází.

Střední hodnota EX náhodného vektoru X je hodnota, kolem které se realizace náhodné veličiny pohybují. Je závislá na typu rozdělení, což je v našem případě diskrétní rozdělení, pro něž se střední hodnota vypočítá jako

$$EX = \sum_{x\in X} P(X = x).$$

Rozptyl je střední kvadratická odchylka, která nám udává, jak moc dané hodnoty kolísají kolem střední hodnoty. Rozptyl je pak definován takto

$$varX = E(X - EX)^2.$$

Pokud můžeme pro dvě náhodné veličiny definovat jejich rozptyly, pak lze zavést kovarianci těchto veličin jako

$$cov(X,Y) = E(X - EX)(Y - EY).$$

Je-li kovariance nulová, pak jsou náhodné veličiny nezávislé (nekorelované). Pokud mají X a Y kladné rozptyly, definujeme korelační koeficient jako

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{varX}\sqrt{varY}}.$$

Korelace nabývá hodnot z $\langle -1, 1\rangle$ a čím více se v absolutní hodnotě blíží 1, tím větší je lineární závislost mezi sledovanými veličinami. Naopak pokud je nulová či blízká nule, pak je závislost bezvýznamná či žádná.

Korelační znaménko pak udává, o jaký typ závislosti se jedná. Pokud je hodnota kladná, jedná se o pozitivní korelaci a hodnoty obou veličin rostou v závislosti jedna na druhé. Pokud je hodnota korelace záporná, pak se jedná o negativní korelaci, potom pokud jedna hodnota roste, druhá klesá. [15]

3.5 Stacionarita

Časové řady mají často, narozdíl od klasických náhodných vektorů, tendenci vzájemné návaznosti mezi hodnotami. V datech tohoto typu je často přítomná absence stacionarity, což je způsobeno například přítomností trendu či sezónní periodicity. Stacionární řada vykazuje stejné vlastnosti nezávisle na čase.

Problém pak nastává při pokusu o modelování regrese nestacionárních procesů. V případě dvou nezávislých časových řad Y_t a X_t může metoda nejmenších čtverců při odhadu regresní přímky dát pro dostatečně velký vzorek dat statisticky významný sklon. Pokud například střední hodnoty dvou řad podléhají nějakému (ne nutně stejnému) trendu, regrese pak porovnává přítomnost trendu a ne změn ve směru vývoje obou procesů, a je proto sporná.

Pokud chceme nestacionární řadu převést na stacionární, nejjednodušším způsobem je její diferenciace, kdy je její hodnotou rozdíl proti hodnotě předchozí.

$$Y_t : \Delta Y_t = Y_t - Y_{t-1}.$$

Takto převedená časová řada je nyní řadou rozdílů v čase. Touto diferenciací je odebrána trendovost dané řady a tím i zajištěna její stacionarita. Pokud je třeba, lze diferencovat i vícekrát, aby tak bylo dané stacionarity dosaženo, ale v běžných případech stačí pouze jednou.[13]

3.6 Autoregresní model

Pro ověření stacionarity lze použít autoregresní model AR(p), který vysvětluje hodnotu Y_t za pomoci p předchozích hodnot. Pro AR(1) tedy platí [17]

$$Y_t = \alpha + \rho Y_{t-1} + \epsilon_t,$$

kdy platí, že Y:

- je stacionární, pokud $|\rho| < 1$
- je nestacionární s jednotkovým kořenem pro $\rho = 1$
- je nestacionární s explozivním vývojem pro $|\rho| > 1$

Obecně pak platí vzorec

$$Y_t = \alpha + \phi Y_{t-1} + \gamma_1 \Delta Y_{t-1} + ... + \gamma_{p1} \Delta Y_{t-p+1} + \sigma_t + \epsilon_t.$$

3.6.1 Dickey-Fuller test

Jedním ze základních algoritmických testů stacionarity je Dickey-Fuller test, který testuje přítomnost jednotkového kořene. Pro jeho ověření uvažuje tři rovnice:

$$\Delta y_t = \gamma y_{t-1} + \epsilon_t,$$

$$\Delta y_t = a_0 + \gamma y_{t-1} + \epsilon_t,$$

$$\Delta y_t = a_0 + \gamma y_{t-1} + a_2 t + \epsilon_t.$$

Rozdíl mezi těmito regresemi tvoří deterministické prvky a_0 a $a_2 t$. Za platnosti nulové hypotézy $\gamma = 0$ první rovnice představuje čistý model náhodné procházky, ve druhé rovnici je navíc obsažena úrovňová konstanta a_0 a třetí rovnice obsahuje navíc a_0 i lineární časový trend $a_2 t$. Chceme testovat γ.

Pro získání příslušných hodnot se dané rovnice odhadují pomocí *metody nejmenších čtverců*. Následně testujeme $H_0 : \gamma = 0$ proti $H_A : \gamma < 0$ a výslednou t-statistiku porovnáme s odpovídající hodnotou z Dickey-Fullerových tabulek a rozhodneme o přijetí či zamítnutí nulové hypotézy.

Tento test má výhodu ve využití OLS při odhadu, ale celkově má tento malou vypovídající sílu, a proto se častěji využívá jeho rozšířená verze, ve které se místo autoregresního procesu prvního řádu používá autoregresní proces řádu p v následující podobě:

$$y_t = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3} + ... + a_{p-2} y_{t-p+2} + a_{p-1} y_{t-p+1} + a_p y_{t-p} + \epsilon_t$$

.

Po aritmetických úpravách následně získáme vzorec v podobě

$$\Delta y_t = a_0 + \gamma y_{t-1} + \sum_{i=2}^{p} \beta_t \Delta y_{t-i+1} + \epsilon_t,$$

kde $\gamma = (1 \sum_{i=1}^{p} a_i)$ a $\beta_i = \sum_{j=i}^{p} a_j$. Původní rovnice se pak nahradí a získáme tak rovnice ve tvaru:

$$\Delta y_t = \gamma y_{t-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \epsilon_t,$$

$$\Delta y_t = a_0 + \gamma y_{t-1} + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \epsilon_t,$$

$$\Delta y_t = a_0 + \gamma y_{t-1} + a_2 t + \sum_{i=2}^{p} \beta_i \Delta y_{t-i+1} + \epsilon_t.$$

Následující postup je podobný s klasickým DF-testem, ale díky použití zpožděných diferencí závislé proměnné y má tato verze testu větší statistickou sílu. Také lze v této verzi testovat jednotkový kořen v AR procesech vyšších řádů. Pořád se ovšem nejedná o naprosto spolehlivý test. [18]

3.6.2 KPSS test

Druhým běžně využívaným testem stacionarity je Kwiatkowski–Phillips–Schmidt–Shin test, který vychází z toho, že Dickey-Fuller často nedokáže vyvrátit nulovou hypotézu, tak označí velké množství řad za nestacionární.

Z toho důvodu modeluje řadu jako součet deterministického trendu, náhodné procházky a stacionární chyby

$$y_t = \xi_t + r_t + \epsilon_t,$$

kde r_t je náhodná procházka:

$$r_t = r_{t-1} + u_t,$$

kde u_t je iid náhodná veličina z $N(0, \sigma_u^2)$ a r_t je úrovňová konstanta. Testujeme nulovou hypotézu trendové stacionarity $u_2 = 0$, tedy že dlouhodobý rozptyl je nulový. Pro testování se používá LM statistika:

$$LM = \sum_{t=1}^{T} \frac{S_t^2}{T^2 \hat{\sigma}^2},$$

kde $S_t = \sum_{i=1}^{t} e_i$.

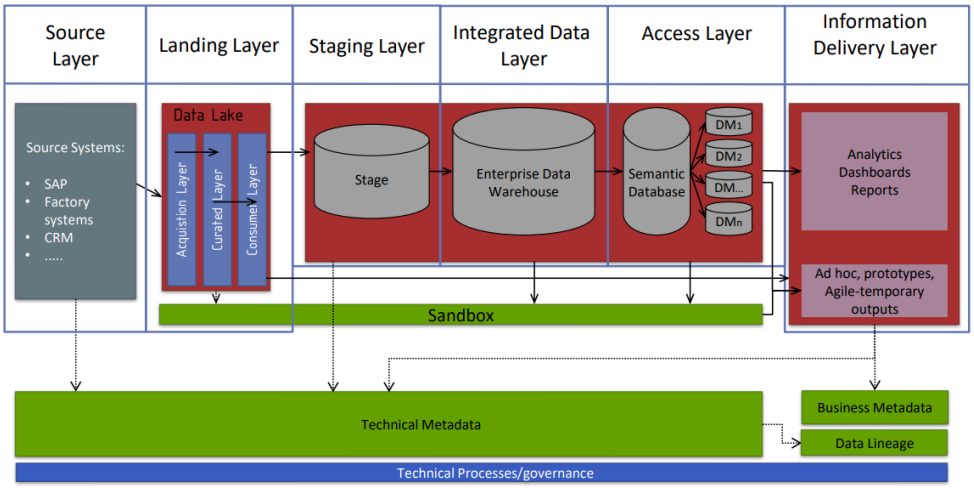
Nulová hypotéza tedy stacionaritu potvrzuje a interpretace p-hodnoty je vůči ADF opačná. Je tedy vhodné pro lepší ověření stacionarity tyto dva testy kombinovat a rozhodovat se na základě výsledků obou testů. [18]

Kapitola 4

Datové sklady

Datový sklad je nejčastěji centralizované úložiště, ve kterém se shromažďují data z různých zdrojů. Tato data se v průběhu procesů, které v datovém skladu probíhají, standardizují, čistí, transofrmují a historizují, aby měla optimální formu pro budoucí zpracování. Tato data jsou následně využita pro různé analýzy, predikce a následné vizualizace. [20]

Source Layer slouží k napojení zdrojových systémů pro automatizaci získávání zdrojových dat. V mém případě jsou zdrojová data nahrána z CSV souborů, které jsou buď ručně získány ze zdrojů, či v případě Google Trends pomocí vytvořeného Python skriptu využívajícího k tomuto učelu sloužící API. Tato vrstva je tedy v mém případě vynechána, neboť majorita využitých datových souboru byla získána ručně. [21]



Obrázek 4.1: Model Datového skladu, přednáška MI-DWH [20]

Landing layer je volitelná vrstva a slouží k případné transformaci dat před samotným nahráním do Stage databáze. V mém případě se jedná zejména o sjednocení názvů zemí, kterého bylo dosažena pomocí dalšího mnou vytvořeného Python skriptu. Dále se jedná pouze o drobné úpravy týkající se špatného formátu datového souboru, které neumožňuje automatické zpracování, jako například vložené řádky textu, který neodpovídá hlavičce, nebo chybějící oddělovače. [21]

Stage Layer obsahuje databázi, do které jsou nahrávány surová data z datových exportů. Slouží zejména pro validaci dat před nahráním do Integrated Data Layer. Tuto část využívám pro kontrolu dat a pro jednodušší následnou manipulaci a čištění pomocí ETL procesů (Extract, Transform, Load) při nahrávání do další vrstvy. Další výhodou je i to, že po nahrání do Stage databáze mají všechna data sjednocený datový typ, což je potřeba pro jejich vzájemné párování a porovnávání. [21]

Integrated Data Layer , jinak také známý jako Target. Do této databáze s jednotným datovým modelem, který popisuje všechny entity zde uložené, se ukládají transformovaná data ze Stage vrstvy. Tato data jsou dle potřeby historizována a jsou zde využity technické sloupce, aby bylo možné jednotlivé záznamy jednoznačně odlišit. Toto je srdce celého datového skladu, ve kterém se všechna potřebná data soustředí, a tedy je tato vrstva implementována i v mém případě. [21]

Access Layer obsahuje jednotlivá datová tržiště, která mají napočítána často agregovaná data na základě specifických požadavků. Často jsou obohacena o sémantickou databází, která zaručuje "jednotnou pravdu"v datových tržištích. V mém případě je vytvořeno datové tržiště a sémantická vrstva je vypuštěna. Jednotná pravda je zde zajištěna pomocí skriptu, který sjednocuje názvy zemí, a také pomocí jednotného formátu datumu. Díky tomu jsou veškeré společné sloupce jednotné. [21]

Information Delivery Layer slouží pro přístup k datům ze strany uživatele. Pro tyto potřeby se využívají různé reportovací, analytické, prediktivní nástroje a modely. Opět pro mé potřeby samozřejmostí a je využito reportovacího systému ve formě dashboardu, který je zprostředkován pomocí Power BI a následně zveřejněn na webových stránkách. [21]

4.1 Extract, Transform, Load (ETL)

Jak již název napovídá, jedná se o tři procesy, které slouží pro kopírování dat z jednoho či více zdrojů do cílového úložiště v jiném kontextu a formě. Tyto procesy se mohou provádět ručně nebo automaticky pomocí k tomu určených programů (Power Query, Pentaho DI,...). Cílem procesů je převést

data do podoby, v které je chceme dále uchovávat. Tato podoba bývá vhodná pro analytické účely a nemusí zde být zajištěna třetí normální forma. Dále pomáhají udržovat čistotu a datovou kvalitu, díky čemuž se ukládají jen data důležitá pro další analýzy a BI.

Někdy se také používají procesy v podobě ELT, kdy se data transformují až po nahrání do finálního úložiště či jejich kombinace v závislosti na situaci. ETL/ELT procesy by měly mít následující vlastnosti: [23]

- Transakční zpracování potřebné pro správnou komunikaci s databázovými systémy a jejich transakcemi (transakce proběhne pouze jako celek, jinak je celá rollbacknutá).
- Kontrola datové kvality aspoň na základní úrovni.
- Zpracování výjimek - musí umět odchytit a vyhodnotit výjimku, která nastane v průběhu zpracování.
- Idempotentnost - opakované provedení procesů vede ke stejnému výstupu.
- Logování - pro přehlednější kontrolu průběhu a zpracování informací při nastalých chybách.

4.2 Historizace dat

Na rozdíl od klasických databází, kde nás zajímá pouze aktuální stav dat, v datových skladech je důležité, i jak data vypadala v historii, aby mohly být jejich změny analyzovány a na jejich základě být vytvářeny predikční a statistické modely.

Protože ne u všech dat je potřeba držet plnou historizaci, tak při práci s datovými sklady existují tři úrovně historizace, které se dnes nejběžněji využívají: [22]

SD0 - Zachovává se pouze stávající hodnota, pokud přijde do databáze pokus o uložení dat pro daný identifikátor a jinými daty, jedná se nejpravděpodobněji o chybu. Operace INSERT.

SD1 - Přepsání stávající hodnoty za novou. Udržují se pouze aktuální data. Operace INSERT, UPDATE.

SD2 - Udržování úplné historie, kdy je pro každý nový záznam vytvořena nová kopie s unikátním technickým klíčem a časovými záznamy doby platnosti. Protože tento typ historizace je nejdůležitější, neboť se udržuje kompletní verzovaná historie, tak pro lepší předsstavu je na obrázku 4.2 ukázka záznamu, kde se pro studenta s ID 2529 a SK 5 změní stav o akademických titulech. Je tedy vytvořen nový záznam s identickým SK

T_ID	SK	id_student	a_academic...	VALID_FROM	VALID_TO	...
3845	5	2529	N/A	1.1.1900	6.4.2015	...
4215	5	2529	Bc.	7.4.2015	1.1.2999	...

Obrázek 4.2: Ukázka záznamu s typem historizace SD2 [22]

a id_student a každému záznamu je přiřazen unikátní T_ID. Dále jsou zde uloženy hodnoty o období platnosti VALID_FROM a VALID_TO.

4.3 Dimenzionální databáze

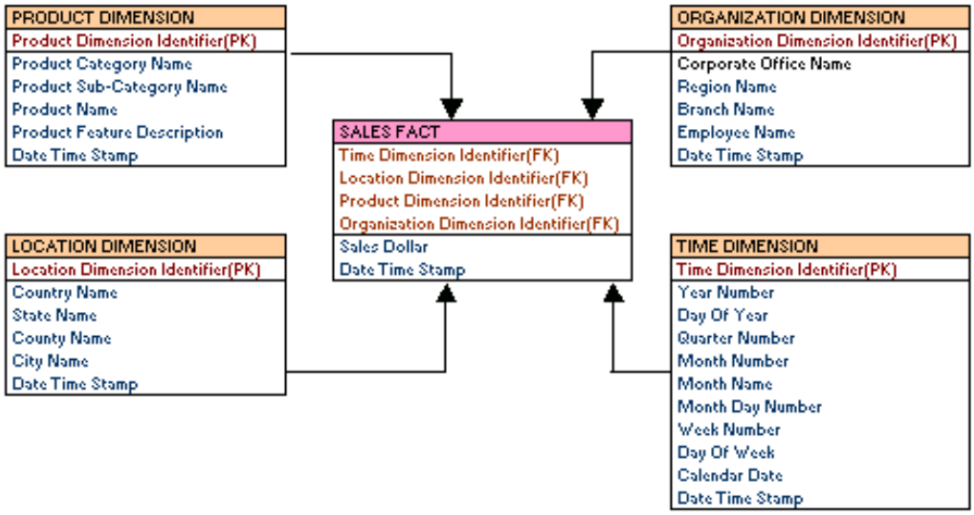
Tyto databáze se nejčastěji využívají v datových skladech v části zvané datamart. Pro jejich implementaci lze využít pouze náhledy do hlavní centrální databáze (targetu) nebo pro ně vytvořit vlastní tabulky. Tyto databáze jsou uzpůsobeny pro potřebnou analýzu daných problémů a každý samotný datamart by měl odpovídat na určitou problematiku či skupině souvisejících problémů. Proto také často obsahují již agregovaná data, která následně urychlují odpovědi na případné dotazy ze strany uživatele.

Díky tomu, že tyto databáze mají různé dimenze náhledu, můžeme zkoumat dané problémy s proměnlivou hloubkou a zaměřením. Nejčastějšími dimenzemi jsou časová, polohová, produktová či kategorická, což souvisí s jejich primárním účelem podpořit business společnosti, jež tyto sklady používá, a je pro ni tedy zásadní zkoumat úspěch jejich prodejen a obchodovaného portfolia.

Nejčastější podoby dimenzionálních tabulek bývají buď hvězda, kdy jsou na faktovou tabulku přímo vázány všechny dimenze, a nebo vločka, kdy mohou dimenzionální tabulky mít podtabulky, do kterých se větví. Zatímco hvězda je jednodušší z hlediska implementace a vzájemného propojení, díky maximální hloubce propojení rovné jedné, tak vločka mnohem lépe podporuje následné úpravy a přidávání dalších kategorií a nové dělení, které vede v případě vločky pouze k připojení nové poddimenze, kdežto u hvězdy je třeba předělat již existující dimenzionální tabulku.

Centrem všeho je faktová tabulka, která je postavena tak, aby odpovídala na určitý dotaz. Tato tabulka na sebe následně váže ostatní dimenzionální tabulky, které ji umožní měnit náhledy na danou problematiku v osách různých dimenzí. Například se můžeme v našem případě podívat na celosvětovou plodnost aktuálně, ale také pouze na plodnost v Evropě či Austrálii, a to v letech 1990 - 2000.

Faktová tabulka dále často obsahuje předpočítané proměnné, které jsou pro dané položené dotazy využívaný, aby se tak šetřil čas, který by jejich okamžitý výpočet ze surových dat trval. [24]



Obrázek 4.3: Příklad schématu dimenzionální databáze v schématu hvězdy. Z cvičení MI-DWH. [24]

4.4 Dashboard

Jedná se o specializované programy, které jsou určeny pro práci s datamarty. Slouží k vizualizaci daných analytických sad a k živé práci s různými dimenzemi u daných faktů. Jsou vhodné pro svou výtečnou prezentaci za použití vhodných vizualizačních prostředků, díky nimž je zajištěna dobrá čitelnost. Pomáhá zodpovídat otázky a podporuje analýzu a rozhodování. Také nám může pomoc utvářen nové pohledy na danou problematiku a pomáhá prezentovat výsledky druhým.

Vizualizace by tedy měla být přehledná a snadno čitelná, aby byly důležité výsledky a hodnoty uživatelům přímo na očích. Tvůrce by se měl vyvarovat složitých a rozptylujících prvků a jiného tzv. šumu. [24]) Nejčastější způsoby vizualizace jsou následující prvky:

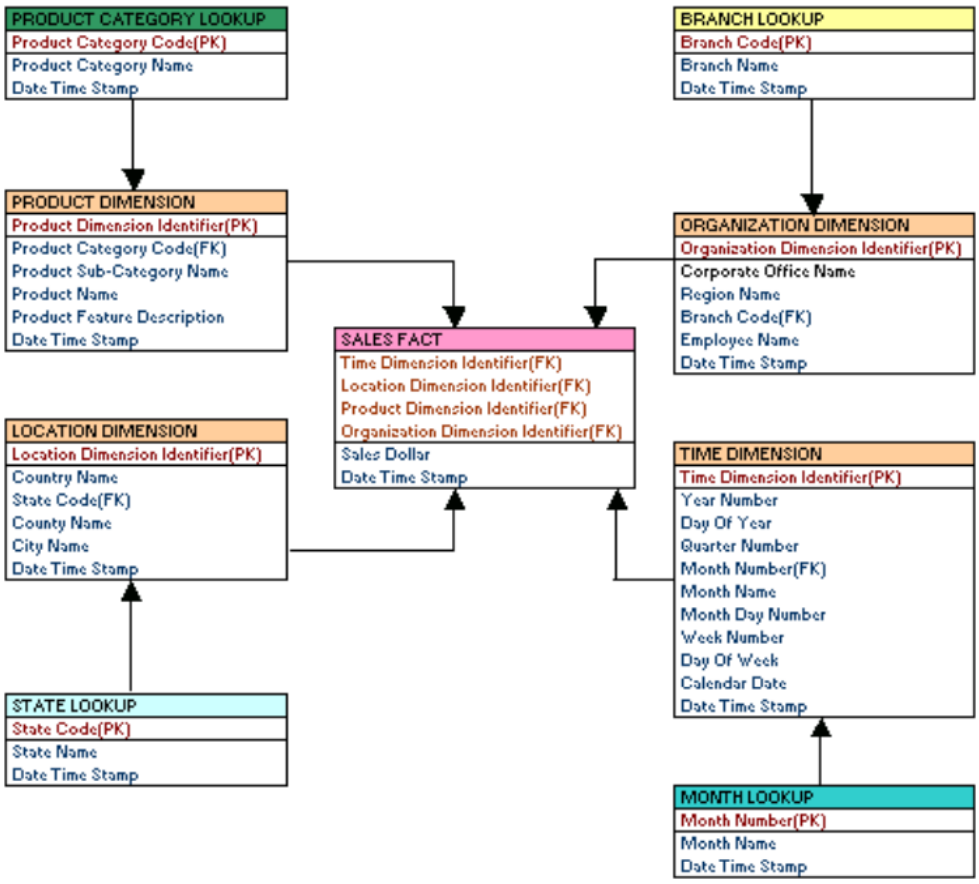
Bar chart - Sloupce jsou ideální pro reprezentaci míry, a jsou tedy vhodné pro porovnávání kategorických dat.

Line chart - Vhodné pro zobrazení změn v průběhu času, pomocí spojnice sledujeme trendy.

Scatter plot - Užitečné pro zobrazení vztahu mezi dvěma proměnnými.

Heatmap - Tabulka, která využívá barvu ke zvýraznění rozdílů mezi jednotlivými hodnotami.

Tabulka - Je vhodná pro zobrazení přesných hodnot. [24]



Obrázek 4.4: Příklad schématu dimenzionální databáze v schématu vločky. Z cvičení MI-DWH. [24]

4.5 Business intelligence

Business intelligence je sada procesů, know-how, aplikací a technologií, které slouží pro optimalizaci procesů, umožňují získat přehled o dění v instituci a vytváří určité výhody oproti konkurenčním institucím.

Jedná se tedy o celkovou disciplínu, která využívá různé technologie a principy a pomocí nich a analytických metod dochází k závěrům, které mohou sloužit pro optimalizaci a vylepšení každodenních procesů. V případě této práce zkoumáme socioekonomické aspekty a vlivy na TFR, kdy po následné analýze různých faktorů můžeme dojit k jisté optimalizaci procesů, jenž s danými faktory souvisí, a tím i přispět k následnému budoucímu růstu úhrnné plodnosti i třeba jen o malé, leč významné hodnoty.

Ve spojení s BI se zkoumané problémy řeší za pomocí KPI (key performance indicator = klíčové ukazatele výkonnosti). Tyto ukazatele nám mají referovat míru vlivu a úspěšnosti v dané zkoumání části. V našem případě

Obrázek 4.5: Dashboard v Power BI. Zdroj prezentace MI-DWH [24]

mohou být takovými hodnotami, které nám ukazují míru vlivu daného faktoru na TFR, například základní statistické hodnoty, jako jsou například míra korelace, či p-hodnota a s ní související výsledky testů hypotéz. [19]

Kapitola 5

Návrh datového skladu

Cílem této kapitoly je seznámit čtenáře s návrhem a podobou datového skladu, který je využit pro uložení, vhodné transformování a následnou analýzu všech vybraných datových sad.

Pro potřeby tvorby všech procesů potřebných k vývoji byly využity nástroje k tomu určeny. Tyto nástroje byly získány v rámci školní licence či v podobě freewaru, který má samozřejmě své nedostatky, ale v rámci této práce je zcela dostačující.

Prvním nástrojem je **Enterprise Architect** [29], který je velmi oblíbený jak při práci během našeho studia, tak i v mnohých společnostech. Tento nástroj je skvělý pro vytváření všech typů modelů a pro tuto práci byl využit pro návrh databázových modelů. Díky jeho funkci generování DDL souborů byly následně z modelů vytvořeny SQL skripty, které mají za úkol připravit v určené databázi všechny potřebné tabulky, do nichž se budou následně automaticky nahrávat veškeré potřebné soubory a data.

Databaze je postavena na **PostgreSQL** [27] a je spravována v rámci vývoje v aplikaci **Postbird** [28], která nabízí všechny služby nutné pro správu všech databází potřebných pro tuto práci.

Všechny automatické ETL procesy byly zpracovány pomocí **Pentaho Data Integration** [30], což je freeware, který je určen pro automatizaci zpracování datových sad. Je skvělý pro načítání surových dat a jejich následné vkládání do databáze a samozřejmě jejich transformaci do podoby vhodné pro následnou analýzu. Protože se jedná o freeware, má samozřejmě i jisté nedostatky, které byly řešeny různými kličkami a případně bylo využito možnosti spouštět i vlastni Python skripty, které jsem pro potřeby další transformace a sjednocení dat vytvořil.

Samotné python skripty jsem vyvíjel ve **Visual Studio Code** [31], což je skvelé IDE pro práci ve všech možných programovacích jazycích. Pro své potřeby jsem s využitím knihovny **pytrends** [33] vytvořil skripty pro automatické stahování dat z Google Trends, který je jedním z mých důležitých datových zdrojů. Dále jsem využil knihovny **country_converter** [34], kte-

rou jsem potřeboval pro sjednocení názvů zemí, aby bylo možné dané sady snadno strojově porovnávat (Neboť "Czechia" a "Czech Republic" je pro SQL dotaz inner join i samotnou relaci problém).

Součástí skriptu je i ověření stacionarity datových sad, k čemuž slouží knihovna **statsmodels** [35], z níž využívám funkce *adfuller* a *kpss*. Pokud je následně u datového souboru potvrzena nestacionarita, jsou jeho hodnoty diferencovány pomocí knihovny **numpy** [36]. V poslední řadě nelze zapomenout na knihovnu **Pandas** [37], která obsahuje nástroje pro práci s Dataframy, jež jsou vhodné pro ukládání těchto statistických dat, a také funkce pro načítání a ukládání csv souborů.

Pro samotnou analýzu dat a jejich vhodnou vizualizaci jsem zvolil nástroj **Power BI** [32], jenž je součástí *Microsoft Office 365*, který máme k dispozici v rámci školní licence. V tomto nástroji je možné vytvářet libovolné vizualizace na datech připojených jak ze souborů, tak z databází. Také nabízí šikovné možnosti filtrování dat, které se zobrazují a díky vlastnímu programovacímu jazyku *Power Query M* lze vytvářet i vlastní míry, pokud si uživatel nevystačí s těmi, které jsou již v rámci nástroje předpřipravený. Také obsahuje možnost vytvářet vlastní vizualizace a výpočty na datech pomocí jazyků *R* a *Python*, čehož bylo také využito pro kompletní zobrazení všech potřebných statistických veličin.

5.1 Sběr dat

Data byla získávána z několika zdrojů ve formátu csv. Všechny tyto soubory obsahovaly pro mou následnou analýzu stěžejní hodnoty, jako je země, rok a hodnota. Přestože ne všechny sady obsahovaly veškerá data pro všechny země či časové úseky, pro samotnou statistickou hodnotu to nemělo zásadní vliv, neboť takové hodnoty nebyly na sebe jednoduše napárovány a nebyly tak v následné analýze zahrnuty. To ovšem nevadí, neboť se používala data ze všech dostupných zemí, a tak byl daný vzorek velmi bohatý.

Jedním z nich byl již výše zmíněný **Google Trends** [38], z kterého byla data stažena pomocí vytvořeného Python skriptu. Z důvodu omezení ze strany Googlu, kdy při větším množství dotazů z jedné IP adresy docházelo k přerušení spojení a konci stahování dané sady, jsem do kódu vložil sleep, který měl tento problém eliminovat, ale ani tak nebylo možné tyto velké datové sady stáhnout naráz. Proto jsem je stahoval po částech, které jsem ukládal do samotných csv souborů, které jsem následné pomocí dalšího mnou vytvořeného skriptu sloučil do jediného souboru.

Google Trends poskytuje svá normalizovaná data o vyhledávání určitých témat a frází ze všech dostupných zemí od roku 2004. Normalizovaná jsou tak, že misto surových hodnot jsou jednotlivé hledání v daných zemích hodnoceny ve škále od jedné do sta. Časově jsou data poskytována za každý měsíc, takže

aby odpovídala ostatním veličinám, které jsou sbírány v roční frekvenci, byla tato data zprůměrována, a tím byla získána finální roční hodnota.

Druhým významným zdrojem je **World Bank Group** [39], která pod sebou sdružuje pět organizací, jež mají za cíl obnovu a rozvoj zejména v rozvojových zemích. Tato skupina vznikla v roce 1944 a postupně se rozšiřovala o další dílčí organizace, které jsou nyní její součástí. Sdružuje 189 členských států a její součástí je i Development Data Group, která spravuje databáze shromažďující statistická data členských zemí. To ve výsledku dává bohatou kolekci dat s dostupnými indikátory pro široké územní i časové spektrum. Díky tomu, že jsou tyto data volně dostupná, jsou také vhodným zdrojem získávání potřebných datových sad.

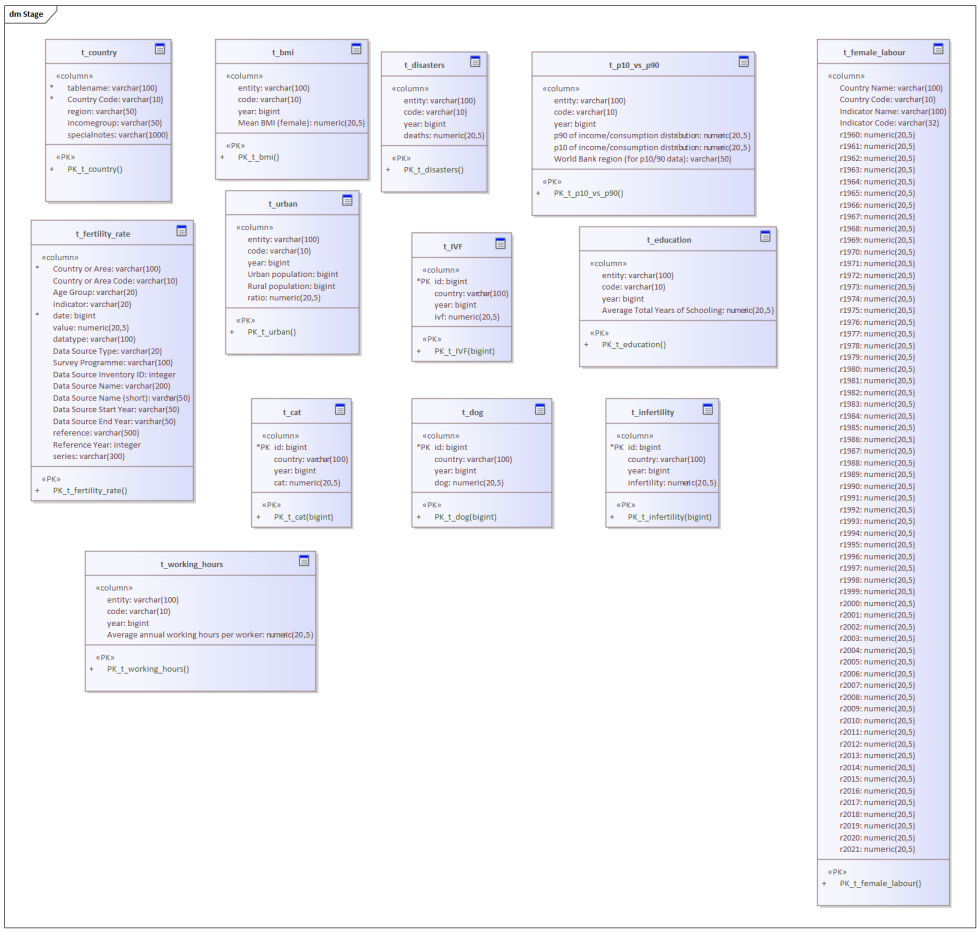
Posledním zdrojem je **Our World in Data** [40], což je výsledek spolupráce mezi University of Oxford a neziskovou organizací *Global Change Data Lab*. Cílem toho spojenectví je zmapovat, jak se v globálním měřítku mění podmínky pro žití, životní prostředí a další důležité ukazatelé, které ovlivňují jak jednotlivé regiony, tak celý svět. K tomu využívá kromě vlastních dat a výzkumů, která jsou vytvořena ve spolupráci s *Oxfordskou Univerzitou* a dalšími dobrovolnými pracovníky, také data a práce, které vytvářejí jednotlivé instituce v daných zemích, a nejsou tak široké veřejnosti dostupné.

Tabulka 5.1: Tabulka datových sad s jejich zdroji a základními parametry dat v jednotlivých sadách.

Název veličiny	**Zdroj**	**Počet zemí**	**Časový rozsah**	**Jednotka**
TFR [49]	Our World in Data	198	1950-2018	num. of kids per woman
Country [47]	World Bank	209	-	country
BMI female [41]	Our World in Data	199	1975-2016	kg/m2
Cat	Google Trends	222	2004-2020	mean year score
Disasters [46]	Our World in Data	203	1990-2019	num. of death per 100 000 inhabits
Dog	Google Trends	186	2004-2020	mean year score
Female labour [48]	World Bank	215	1960-2021	%
Infertility	Google Trends	133	2004-2020	mean year score
IVF	Google Trends	155	2004-2020	mean year score
P90vsP10 [42]	Our World in Data	161	1981-2017	-
Schooling [43]	Our World in Data	192	1870-2017	years
Urbanization [44]	Our World in Data	211	1960-2020	-
Working hours [45]	Our World in Data	70	1870-2017	h

5.2 Stage

Do první databázové části se nahrávají data v surové podobě. V tomto případě se jedná o soubory v podobě, v které byly staženy. V případě dat získaných pomocí skriptu z Google Trends, jenž byl za tímto účelem vytvořen, jsou již



Obrázek 5.1: Obrázek modelu Stage databáze v programu Enterprise Architect.

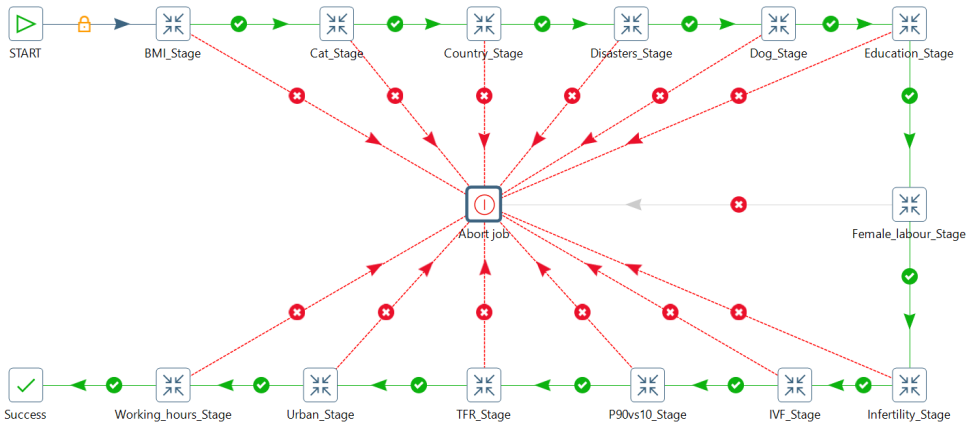
data stahována v potřebné podobě a při slučování dílčích souborů jsou zároveň již sjednoceny názvy zemí.

Databáze se před každým nahráním vyčistí, aby byla udržována vždy jen aktuální iterace a samotná data se zbytečně neduplikovala. Pro každý soubor je vytvořena samostatná tabulka, která obsahuje všechny hodnoty tak, jak jsou v nahraném souboru.

Dále se zde nachází pomocné tabulky, kde se uchovávají dočasná data pro další zpracování pomocí skriptu či následné nahrání do Target databáze po transformaci.

Logický model byl vytvořen v Enterprise Architectu a je tvořen pro PostgreSQL databázi. Pro ni je i generovaný samotný kód pro vytvoření databáze a tabulek, jež obsahuje, a to přímo z již výše zmíněného nástroje.

Když je databáze správně vytvořená a spárována s nástrojem Pentaho



Obrázek 5.2: Pentaho Stage Job se všemi transakcemi, které se postupně dle modelu spouští.

Data Integration, lze pak soubory nahrát automaticky spuštěním Jobu Stage. Ten vždy dané tabulky vyčistí a následně tam nahraje nová data.

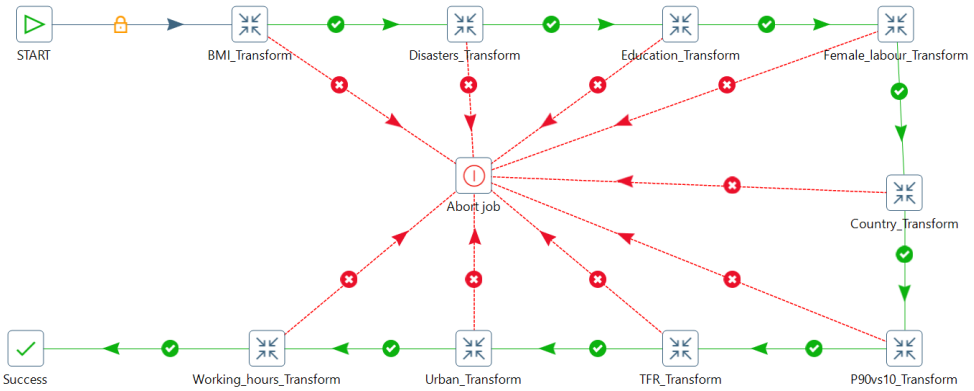
5.3 Target

Databáze Targetu je srdcem celého datového skladu, zde se ukládá a historizuje vše důležité. Proto v průběhu ETL procesů dochází k čištění a transformaci dat tak, aby byla následná práce s nimi co nejefektivnější. Proto zde bylo spoustu sloupců zahozeno a nebo transformováno do jiné, pro naši analýzu vhodnější, podoby. Všechny tyto transformační procesy jsou, při správném nastavení databází a cest v proměnných prostředí v nástroji Pentaho DI, zcela automatické a proběhnou při tvorbě samotného datového skladu.

Všechny transakce, které slouží k předpřipravení všech dat pro transformaci pomocí python skriptu, jsou součástí Pentaho Jobu Transform. Ten data pročistí a případně upraví do vhodné podoby, aby na nich následně mohl být spuštěn daný skript, který má za úkol sjednotit názvy všech zemí a ověřit stacionaritu dat v těchto časových řadách. Pokud jsou data nestacionární, je na nich následně provedena diferenciace, aby byla stacionarita zajištěna. Veškeré mezikroky v transformaci se ukládají do CSV souborů v příslušných složkách, díky čemuž jsou veškeré mezikroky uloženy a je tedy možné sledovat a ověřovat dané postupy.

Při transformaci byly ke každé tabulce přidaný čtyři technické sloupce, které jsou potřebné pro správnou historizaci dat. Jedná se o:

- TK - technický klíč, který funguje jako nový umělý identifikátor
- version - tento sloupec nám udává, o kolikátou verzi dané tabulky v rámci historizace se jedná



Obrázek 5.3: Pentaho Transform Job se všemi transakcemi, které se postupně dle modelu spouští.

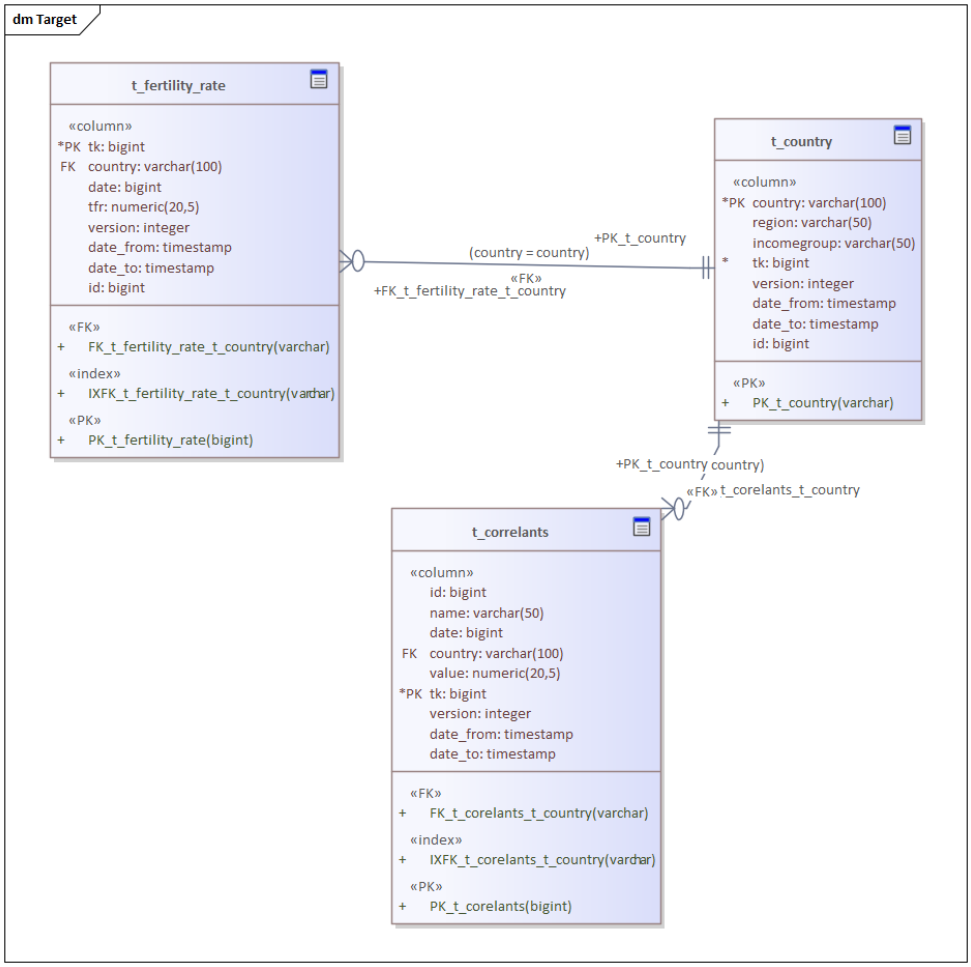
- Date_from - timestamp, který udává, od jakého data je daná verze tabulky platná
- Date_to - timestamp, který udává, do kdy je daná verze tabulky platná

Samotná databáze byla oproti Stage zjednodušená na pouhé tři tabulky, kdy jedna je určena pro TFR (t_fertility_rate). Ta obsahuje následující sloupce vyjma těch technických:

- Country - pro polohovou lokalizaci dané veličiny. Je v relaci s tabulkou t_country, díky čemuž lze hodnoty filtrovat dále dle regionů a příjmů
- Date - pro určení roku, pro který je daná hodnota platná
- tfr - stacionární (diferencovaná) hodnota TFR pro danou zemi v daný rok

Další tabulkou je pomocná tabulka t_country, díky níž lze lépe pracovat s polohovým určením a kromě jednotlivých států lze dané země seskupovat i podle regionů a příjmových skupin, což může pomoc lépe sledovat jejich vzájemné vlastnosti pro určité veličiny.

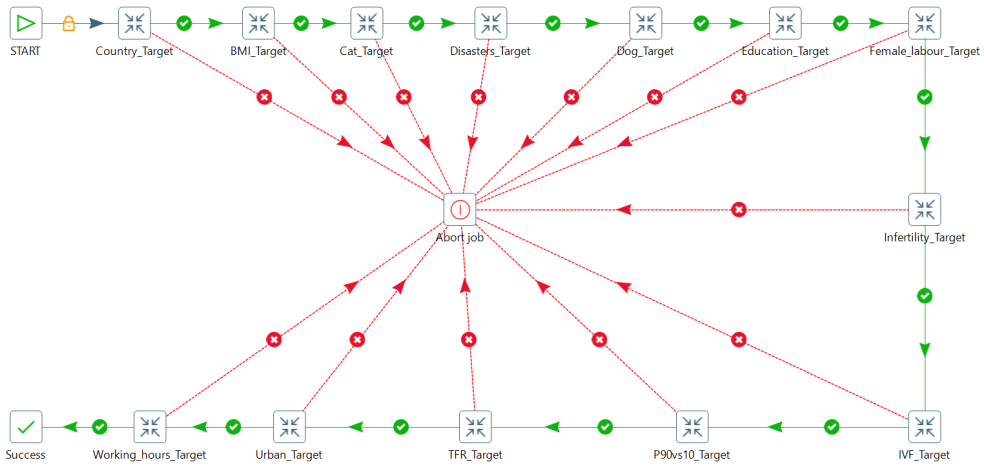
- Country - název země
- Region - jednotlivé země jsou sjednoceny do větších regionů, které mají krom geografie často společné i další vlastnosti
- Incomegroup - rozdělení zemí podle jejich ekonomického hodnocení a výkonnosti



Obrázek 5.4: Obrázek modelu Target databáze v programu Enterprise Architect.

Poslední tabulka t_correlants obsahuje všechny ostatní veličiny, které chceme porovnávat vůči TFR, a má následující strukturu:

- Country - stejně jako u TFR je vázána na t_country a spolu s datem slouží k párování veličin
- Date - rok, pro který je daná veličina měřena
- Name - název veličiny, kterou chceme porovnávat s TFR. Je určen pro následné filtrování jednotlivých veličin v této tabulce
- Value - obsahuje samotnou hodnotu dané veličiny, u které je také po transformaci zajištěna stacionarita



Obrázek 5.5: Pentaho Target Job se všemi transakcemi, které se postupně dle modelu spouští.

I tato část má v Pentaho vytvořený Job s názvem Target, který slouží k nahrání dat z pomocných souborů. Dále také vytvoření všechny pomocné sloupce a nahraje data do Target databáze v historizované podobě.

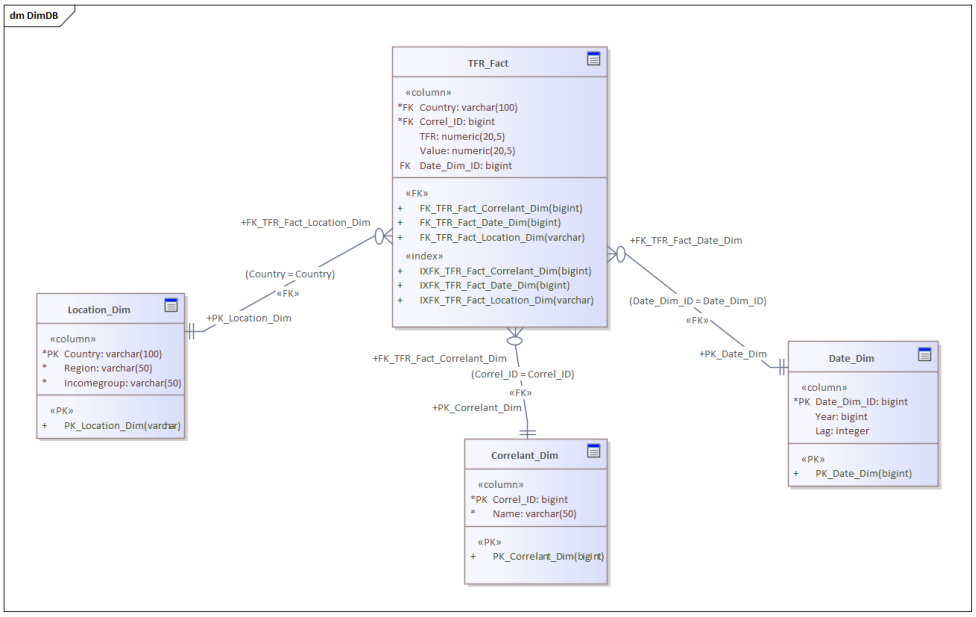
5.4 Data Mart

Tabulky dimenzionální databáze jsou vytvořeny pouze pohledy na samotnou Target databázi. Tím se ušetří prostor a pro potřeby naší analýzy je to zcela dostačující. Problém by nastal pouze ve chvíli, kdy by na tyto pohledy přicházelo v jednu chvíli velké množství dotazů a nebo by se původní tabulky neustále měnily, ale to není náš případ.

Zde je vidět kompletní model dimenzionální databáze, kde hlavní tabulka krom odkazů na své dimenze obsahuje také vlastní hodnoty veličin, které jsou potřebné pro přímou analýzu. Jak je z modelu na obrázku 5.6 patrné, dimenze korelantů a datumu jsou velmi triviální, a tak je pro zjednodušení vhodné je při reálné tvorbě pohledů sloučit se samotnou tabulkou faktů. Toto sloučení je patrné z následujícího obrázku 5.7, který představuje již model v takové podobě, v jaké se nachází v dashboardu Power BI a z kterého se vytváří samotné vizuály.

Základní parametry, přes které je možné následně v samotném vizuálu dané hodnoty filtrovat, jsou v Location_dim:

- Country - lze sledovat hodnoty pro určitou zemi
- Region - vybere skupinu zemí daného regionu, v kombinaci s volbou Country lze celkový výběr následně upravit



Obrázek 5.6: Obrázek modelu dimenzionální databáze v programu Enterprise Architect.

- Incomegroup - omezí země na danou skupinu dle jejich ekonomického hodnocení

Dále v Date_dim jsou tyto dva parametry:

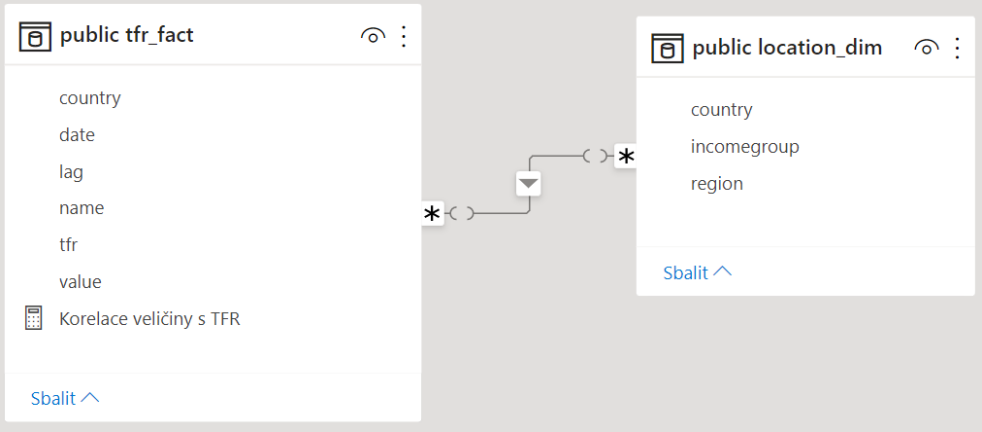
- Date - pro určení přesného roku či časového úseku, pro který se data testují
- Lag - pro určení časového zpoždění sledované veličiny vůči hodnotě TFR

Posledním parametrem je pak **Name**, který nám umožňuje vybrat proměnnou, již chceme aktuálně porovnávat s TFR.

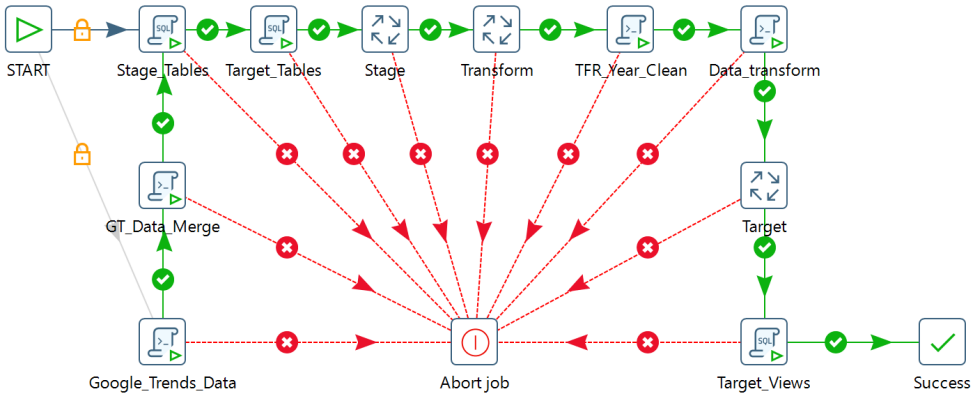
Zjednodušený reálný model již obsahuje pouze dvě tabulky a veličiny z dimenzionálních tabulek jsou nyní součástí faktové tabulky. Dále zde mám nastavenou novou míru, která počítá korelaci mezi zvolenou veličinou a TFR. Ta slouží pro okamžité zobrazení dané hodnoty, neboť zbylé statistické hodnoty jsou počítány pomocí python skriptu, který je přímo součástí vizuálu, ale zde dochází často k částečnému zpoždění, než jsou dané hodnoty zobrazeny.

5.5 Vytvoření datového skladu

Pro vytvoření kompletního datového skladu se všemi daty na vlastní straně je potřeba mít nainstalovanou PostgreSQL databázi, Pentaho Data Integration,



Obrázek 5.7: Obrázek dimenzionální databáze, s kterou pracuje Power BI.



Obrázek 5.8: Pentaho Job pro vytvoření datového skladu. Spouští k tomu postupně všechny potřebné skripty a Pentaho Joby.

Python 3.0+ s příslušnými knihovnami zmíněnými v kapitole Návrh datového skladu a pro následnou analýzu za pomocí dashboardu Power BI.

Kompletní návod je součástí souboru Manual v přiložené složce. Všechny datové soubory jsou také již součástí složky a není tedy třeba je stahovat znovu. Po vytvoření všech potřebných komponent a nastavení všech cest a proměnných dle manuálu lze následně celý sklad automaticky vytvořit spuštěním Pentaho Jobu s názvem *Create_DWH*, jehož struktura je na následujícím obrázku.

Součástí Jobu jsou i skripty pro stahování datových sad z Google Trends a jejich následné spojení do jednoho souboru. Tato část je primárně odpojena, neboť při stahování většího objemu dat je toto API chráněno proti přetížení, a proto tento skript obsahuje sleep, který má minimalizovat šanci na nucené

ukončení ze strany Googlu. Díky tomu ovšem tento skript může běžet dlouhé hodiny a i přes obsažené čekání může server vyhodit chybu a proces ukončit. Tyto data jsou ovšem již stažena a skript lze využít pro samotné stahování nových datových sad.

Když je kompletně vytvořen celý datový sklad (Create_DWH job skončil úspěchem), lze následně na přiložený Dashboard připojit vytvořené náhledy a vložit tak do vizuálů aktuální data a začít vlastní analýzu.

Při vkládání vlastních nových datových sad je třeba vytvořit pro tyto sady nové Transakce v Pentaho DI a pro kompletní vytvoření je vhodně zařadit do stromové struktury. Při ukládání nových dat do Target databáze je třeba dodržet danou datovou strukturu a celý soubor vhodně transformovat.

Kapitola 6

Analýza a zobrazení výsledků

V této kapitole se budu věnovat převážně analýze a interpretaci výsledků, ke kterým jsem pomocí nástrojů BI došel. Budu zde blíže rozebírat jejich vizualizaci a význam jednotlivých prvků zobrazení, aby mohl uživatel sám zkoumat jednotlivé hodnoty, které toto zobrazení nabízí. K tomu, aby byly tyto výsledky poskytnuty co nejširšímu množství uživatelů, je jejich interaktivní dashboard publikován na webové stránky.

V poslední části se budu věnovat vlastní interpretaci výsledků, ve které shrnu celkové výsledky pro vybrané veličiny a zhodnotím možné silné korelanty. Následně se budu věnovat bližší analýze zajímavých veličin, které sice nemusí z celkových výsledků dát silnou korelaci a být tak velký ovlivňovatel TFR, ale existuje u nich nějaký výrazný trend, který je zajímavý a při jeho vhodném aplikování může s dalšími faktory pomoci zvýšit porodnost, ať celkově nebo v určitých zemích, s jejíž kulturou je tento fenomén spojen.

6.1 POWER BI

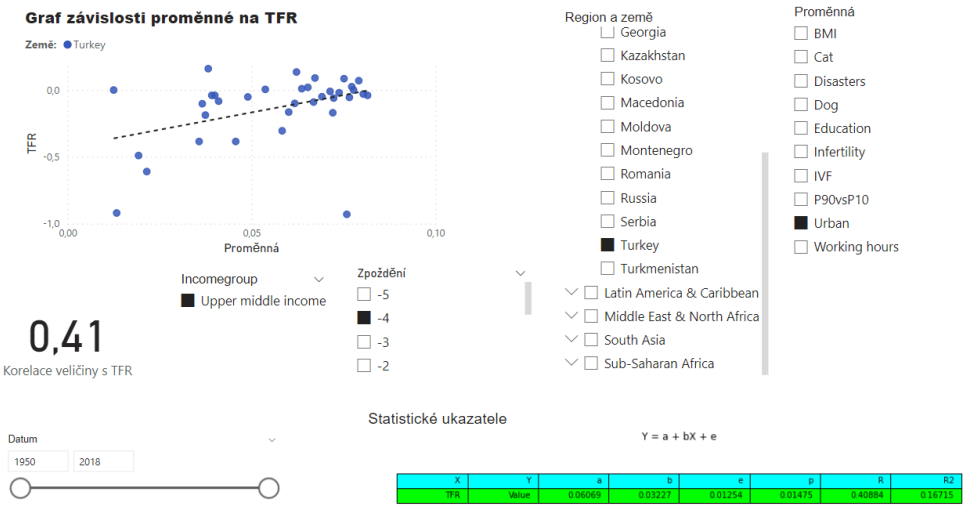
Jedná se o interaktivní software pro vizualizaci dat, primárně zaměřený na business intelligence. Je vyvíjen společností Microsoft od roku 2011 a je součástí Microsoft Power Platform. Podporuje přímé datové vstupy z databází, souborů (CSV, XML, JSON, spreadsheets,...) nebo webových stránek.

Tato aplikace existuje jak v Desktopové verzi, tak v Cloudové, kde funguje metodikou Software as a service. Dále obsahuje další komponenty, které slouží k připojení dalších periferií, jako jsou mobilní telefony či jiné služby, díky čemuž má uživatel neustálý přístup ke všem svým vizuálům.

V této aplikaci jsem prováděl hlavní část analýzy faktorů ovlivňujících TFR. Využil jsem jejích bohatých možností pro připojení dat a jejich vzájemné propojení pomocí vizuálního modelu jednotlivých tabulek. Poté jsem využil valstní tvorby dashboardů, které jsem sestavoval z široké palety vizuálních prostředů a filtrů, které tato aplikace nabízí.

6.2 Dashboard

Dashboard je vytvořen, tak aby dával co největší možnost při nastavování různých kritérií, pro něž se následně vypočítávají potřebné statistické ukazatele.



Obrázek 6.1: Ukázka mnou navrženého dashboardu v Power BI.

Hlavní součástí je samotný graf, který zobrazuje závislost proměnné na TFR, a všechny vybrané státy jsou barevně odlišeny. V grafu je také proložená přímka, která značí vizualizuje lineární regresi pro dané hodnoty.

Dále je tu tabulka se statistickými ukazateli, která zobrazí jak základní parametry pro danou regresní přímku, tak statistické ukazatele, kterými jsou p-hodnota, korelační koeficient R a deterministický koeficient R^2. Zároveň tato tabulka mění barvu podle toho, jestli je p-hodnota vyšší než 5% hranice, pak daná tabulka zčervená, a nebo nižší, kdy tabulka změní barvu na zelenou. Protože je pro tuto část vizuálu použit vložený python script, může jeho zobrazení kratší dobu trvat.

Také je zde samotný ukazatel korelace, který již zobrazuje hodnotu bez zpoždění, a je tak možné sledovat pro rychlé hledání pouze tuto hodnotu a nebo ji využít pro kontrolu hodnot v tabulce (když se shoduje s hodnotou R, pak je tabulka již aktualizovaná).

Kromě těchto ukazatelů jsou zde všechny potřebné filtry hodnot, jedním z nich je časový, který je nastaven ve formátu od-do, ale lze přepnout i do formátu výběru pro zkoumání vývoje hodnoty v různých zemích v daném roce. Dalším filtr umožňuje vybírat země dle jejich příjmu a tím sjednotit ekonomicky podobně vyspělé skupiny.

Další filtru umožňuje dané testované veličiny vůči sobě časově opožďovat, díky čemuž lze najít nějaký ideální trend v tom, jak rychlý má daná proměnná vliv na TFR či naopak. Asi nejdůležitějším filtrem je samotný výběr proměnné,

kterou chceme s hodnotami TFR porovnávat. Jejich názvy koresponduji s názvy u sloupce name v databázi.

Poslední je možnost filtrování regionů a zemí, což kromě přidaného vizuálu umožňuje přímo i graf, kde jsou jednotlivé země umístěny v horní části grafu a při výběru jedné z nich je daná země v grafu zvýrazněná a hodnoty u ukazatelů se přepočítají pro tuto zemi.

6.3 Webové stránky

Osobně si myslím, že jednou z možností, jak TFR zvýšit, je i rozšiřovat povědomí o tomto problému mezi lidmi, aby si uvědomili, že mít děti je důležité. Z tohoto důvodu chci, kromě samotného zkoumání různých korelantů, dát širokému okolí možnost si sami tyto vazby a vztahy prohlédnout. Toho mám v plánu docílit založením webových stránek, které budou obsahovat můj interaktivní dashboard, a dá jim tak možnost nahlédnout do vizualizací a prostých statistických čísel, které dávají různé veličiny ve vztahu k TRF.

Samotný dashboard je hlavním obsahem těchto stránek, a tak se zde budu věnovat převážně jemu a jak je možné publikovat tyto interaktivní analytické vizuály na stránky přímo z Power BI. Celý svůj dashboard jsem vytvořil v Power BI Desktop, ale ten tuto možnost nemá, je třeba ho nejdříve publikovat do svého online pracovního prostoru, který je součástí balíku Microsoft Office 365. Tato verze má sice, co se vizualizačních a analytických možností týče, trochu menší paletu, ale je tu možnost sdílet tento pracovní dashboard na webové stránky.

Pro tuto možnost lze vygenerovat kód, který lze vložit do přímo do kódu stránky, a po jejím načtení se načte i samotný dashboard, pokud je sdílený či originální dashboard aktivní. Pro plynulost práce s dashboardem a zajištění dostupnosti i při větší vytíženosti ze strany koncových uživatelů se dashboard na klientské strane cachuje. Z tohoto důvodu se všechny změny v originální verzi okamžitě nepropisují a je třeba je buď propsat nuceným kompletním načtením dané stránky a nebo po uplynutí TTL (Time to live) dané cache, která se následné aktualizuje sama.

Samotný dashboard funguje v tomto zobrazení identicky jako v samotné aplikaci. Uživatel zde může libovolně změnit zobrazované hodnoty pomocí filtrů, které jsou zde pro všechny primární dimenze. Vizuály, které zobrazují buď číselné, nebo grafické výsledky, se tomuto výběru interaktivně přizpůsobují a při najetí myší na určitý bod v grafu lze i zobrazit detail pro danou hodnotu (rok, hodnota, země).

Samotné stránky jsou aktuálně vytvořeny na mém Google účtu a od kaz na ně je zde:

`https://sites.google.com/view/totalfertilityrate/domovsk%C3%A1-str%C3%A1nka`

V jejich stránky jsem se snažil držet minimalismu, a proto je zde ustředním faktorem právě daný dashboard a krom toho jen úvodní odstavec, který spojuje dané stránky s touto prací a následně pár vět, které návštěvníkům přiblíží problematiku TFR.

6.4 Interpretace výsledků

V této kapitole bych rád shrnul výsledky pro všechny veličiny, které jsem vůči TFR porovnával a u kterých jsem se domníval, že by mohly mít na tuto veličiny nějaký významný vliv, ať už kladný, nebo záporný. Od předchozích kapitol a hlavně od samotného dashboardu se toto shrnutí liší v tom, že obsahuje komplexní výsledky pro všechny hodnoty, a ukazuje tak, jak si dané veličiny v porovnání s TRF vedou napříč všemi roky a zeměmi, pro které byla data vyhodnocována. Na druhou stranu dashboard dává, více než tento ucelený pohled, možnost uživateli zkoumat dané vztahy více do hloubky, díky všem možným filtrům, díky kterým lze nastavovat parametry všemi možnými směry z pohledu dimenzionální databáze.

Pro výpočet této statistiky jsem opět vytvořil krátký program v pythonu (stat_sum.py), který vezme data, jež jsou uložena v databázi v podobě pohledu tfr_fact, a který tyto data pro každou veličinu zvlášť statisticky porovnájí s veličinou TFR. Tento výpočet se provádí za pomocí knihovny scipy, jež testuje lineární regresi mezi danými hodnotami pro každou zemi. Ty země, kde se s pravděpodobností na 95 procent potvrdí, že tyto veličiny korelují, se přičtou do počtu statisticky potvrzených ($p <= 0.05$), a tím se získá celkový počet zemí, kde je tato veličina statisticky významná. Dále se toto provádí pro časové zpoždění veličiny o -5 až 5 let a vybere se taková hodnota zpoždění, která má nejlepší statistické hodnoty (je významná pro největší počet zemí). Díky této veličině pak lze i vysledovat, kdy lze očekávat nejvýznamnější reakci druhé veličiny při určité výrazné změně té první.

Výsledky, které jsem tímto výpočtem získal, jsou sumarizovány v tabulce 6.1. Pro každou testovanou veličinu jsou v ní obsaženy tyto hodnoty popořadě:

- Celkový počet zemí, pro které se daná veličina testovala.
- Celkový počet zemí, pro které je lineární regrese potvrzena na hladině pravděpodobnosti 95%.
- Počet zemí, pro něž je korelace kladná (pro ověření, jestli daná veličina ovlivňuje TFR pozitivně, nebo negativně).
- Nejnižší nalezená hodnota korelačního koeficientu pro danou veličinu, pro níž vyšla lineární regrese statisticky významná.
- Nejvyšší nalezená hodnota korelačního koeficientu pro danou veličinu, pro níž vyšla lineární regrese statisticky významná.

- Pro jaké zpoždění vyšly nejlepší statistické výsledky.

Tabulka 6.1: Tabulka s kompletními výsledky korelace jednotlivých veličin vůči TFR

Název veličiny	Počet zemí	p <= 0.05	Kladá korelace	Nejnižší korelace	Nejvyšší korelace	Zpoždění
BMI u žen	163	37	19	-0.784	0.704	-4
Doba vzdělání	141	29	3	-0.679	0.49	-3
Urbanizace	172	26	11	-0.76	0.56	5
Kočky	113	16	4	-0.79	0.537	-3
Ženská zaměstnanost	175	14	7	-0.525	0.577	-1
Katastrofy	146	11	5	-0.65	0.646	1
Psi	70	10	3	-0.77	0.603	-4
Neplodnost	67	9	4	-0.684	0.63	-4
Pracovní doba	65	8	3	-0.484	0.368	-1
Umělé oplodnění	72	8	4	-0.59	0.764	-5
P90vsP10	26	5	3	-0.47	0.545	-4

Nejlepší výsledky vyšly pro veličinu BMI u žen, která má kromě největšího množství statisticky významných výsledků také velmi vysoké maximum a minimum korelačního koeficientu. Další zajímavé pozorování o této veličině následně shrnuji v samostatné kapitole.

Další významné výsledky vychází pro veličinu Průměrné doby vzdělání, která je již mnoha výzkumy potvrzena jako významný ovlivňovatel hodnoty TFR, a tak je díky tomu zde prokázáno, že moje postupy dávají pro tuto hodnotu očekávané výsledky. Přesto chci tuto veličinu blíže shrnout v samostatné kapitole, neboť se domnívám, že z bližšího zkoumání lze vytěžit zajímavé poznatky.

V samostatné kapitole chci také shrnout třetí nejvýznamnější korelant, kterým je urbanizace, a také trochu blíže rozvést zjištěné výsledky pro domácí mazlíčky (psi a kočky), neboť i u nich dávalo bližší zkoumání poměrně zajímavé výsledky.

Ostatní veličiny již samy o sobě nedávají tak silné výsledky, ovšem důvodů může být více. Například ženská zaměstnanost, která byla velice dobře zkoumána v práci *Influence of women's workforce participation and pensions on total fertility rate: a theoretical and econometric study* [8], je velmi ovlivněna časovým obdobím, kdy je její vztah zkoumán, a proto v tomto celkovém pohledu nemá tak silně vypovídající váhu. stejně tak to může platit i pro další hodnoty a je třeba pak pro lepší pochopení použit například mnou vytvořený dashboard, ve kterém lze libovolně volit sledované období a i z grafu lze snadno vyčíst, které hodnoty v kterém roce se značně odchylují od zbytku.

V poslední řadě, než se dostaneme k zajímavému shrnutí významných hodnot, lze ještě podotknout, že dalším důležitým faktorem je také to, že nikdy tyto veličiny nepůsobí na TFR jen samy o sobě, a tedy v kombinaci s dalšími hodnotami mohou dávat lepší ucelený výsledek a dostat tak i významnější výsledky. Příkladem by mohla být například kombinace neplodnosti a umě-

lého oplodnění, která v kombinaci s vyhledáváním mohou mít významnou souvislost.

6.4.1 BMI

Jak již bylo řečeno, BMI má, co se týče lineární regrese ve vztahu s TFR, velice dobré výsledky, ale je tu pár věcí, které z těchto strohých výsledků zcela nevyplývají. Asi většina z nás ví, že BMI je tabulkový ukazatel "zdraví", lépe řečeno poměr váhy ku výšce. Já osobně těmto hodnotám příliš velký význam nepřikládám, protože mnohem důležitější je složení této váhy. Můžeme to demonstrovat na příkladu dvou osob, kdy jedna má BMI optimální, ale přesto trpí takzvanou skrytou obezitou, protože má minimum svalové hmoty a velké množství tuků, zatímco druhá osoba může mít dle BMI nadváhu, ale přitom má klidně méně než 10 % tělesného tuku a jedná se o silového sportovce s velkým množstvím svalové hmoty.

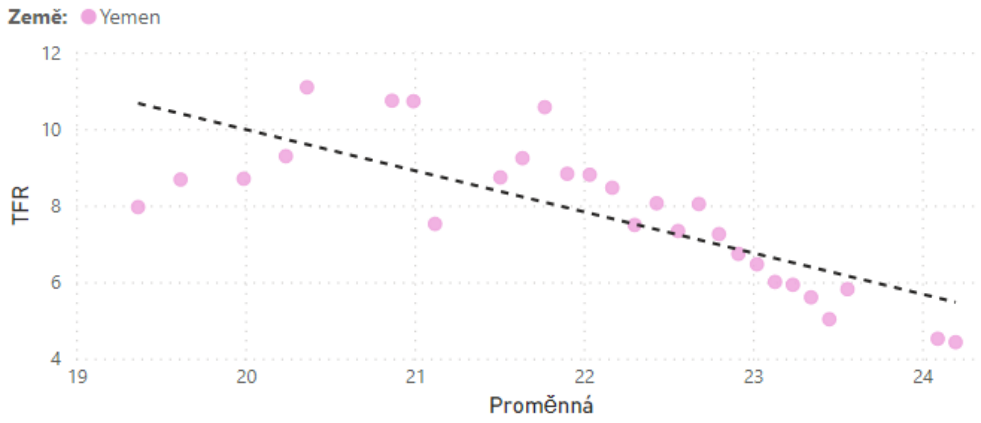
Ale abych se dostal k tomu hlavnímu, ačkoli BMI není optimální ukazatel, pro průměrný vzorek populace poslouží poměrně rozumně. A s tím souvisí právě vztah k TFR, jak je vidět při porovnání surových hodnot BMI a TFR, nejlepší výsledky pro porodnost jsou právě v ideálních hodnotách BMI. I proto, jak je v tabulce 6.1 vidět, je počet kladných a zapornách korelací vyrovnaná, neboť v některých státech se hodnota BMI pohybuje od nízké k normální, a má tedy korelaci kladnou, a v jiných naopak od normální k vysoké, a pak se to láme a korelace padá k záporným hodnotám.

Má to hned několik důležitých důvodů. Jedním velmi známým je, že jakýkoliv extrém (velmi nízká hladina tuku i vysoká obezita) vede k přerušení ženských reprodukčních schopností a neplodnosti, i když ne trvalé. Dále s tím souvisí i fyzická přitažlivost a samotná psychická pohoda. Myslím si, že je určitě mnoho dalších faktorů, které se v tomto mohou promítnout, a proto beru tento výsledek za velmi zásadní.

6.4.2 Vzdělání

Jak již bylo popsáno, vzdělání je klíčový faktor, který TFR ovlivňuje, hlavně z důvodu odkladu prvního dítěte na konec studia a ztráty tak drahocenných let, kdy je žena nejplodnější, ale také proto, že vyšší studium otevírá lepší možnosti, jak se pracovně uplatnit a budovat slibnou kariéru.

Z těchto důvodů, a jak je i z výsledků v tabulce 6.1 zřejmé, je právě doba vzdělání v korelaci s TFR záporná, a tak velmi silně ovlivňuje její pokles. Ovšem co je v tomto případě důležité, je to, že tam funguje určitý vývoj, který přesně kopíruje vliv vývoje od zemí třetího světa až po země vyspělé. Kdy z počátků kdy začne doba vzdělání růst, tak TFR vede k rychlému pádu, ale s postupem času začne brzdit a s každým dalším rokem k době studia je již daný vliv menší a menší.



Obrázek 6.2: Příklad datové sady, kde je vidět maximum hodnot mezi hodnotami BMI 18 a 25. Když se blížíme k těmto mezním hodnotám, tak hodnoty TFR významně klesají. Jemen byl vybrán, neboť obsahuje obě mezní hodnoty. Jiné země se většinou pohybují pouze v horní nebo dolní části křivky.

Jak jsem již zmínil myslím si, že tento faktor velmi silně souvisí s celkovým přerodem v dané zemi a že odpírat ženám vzdělání určitě není správné řešení. Spíše by mohlo pomoci jim již během studia tuto problematiku osvětlit, aby věděly, jak je důležité mít děti.

6.4.3 Urbanizace

Stejně jako vzdělání, tak i tento faktor silně souvisí s postupným přerodem zemí. I tato veličina má velmi dobré výsledky, a je tedy významným korelantem, ale na druhou stranu je z výsledků vidět, že významná část korelantů nabírá kladnou hodnotu. A také to, že na rozdíl od ostatních je doba zpoždění kladná, což znamená, že TFR reaguje se značným zpožděním na rostoucí míru urbanizace.

Tyto dvě zásadní věci souvisí hodně na typu urbanizace, která v dané zemi probíhá, protože pokud je to rychlá a neřízená urbanizace, je výsledek korelace většinou záporný a výsledek není nijak silně ovlivněn zpožděním. Hlavním důvodem je právě to, že tato urbanice není řízená, a není tak v daných městech dostatečně rozvinuté zázemí pro děti a lidé jsou mnohem více nuceni pracovat, aby se v městě, které je dražší než venkov, uživili.

Na druhou stranu je právě pozvolnější přerod, kdy jsou již města na nárůst obyvatel připravena, mají dostatek školek, škol a dalších institucí, kde je o děti postaráno, a tak když se lidé do města nastěhují a po počáteční době, než se zabydlí, stabilizují své příjmy (což je daný lag), se rozhodnou děti mít, tak je výsledná korelace pozitivní, a jsou to ty hodnoty, které jsou ke statistickým výsledkům přidány až díky zpoždění, neboť bez něj tyto hodnoty nebyly statisticky průkazné.

Toto zjištění může být návodné pro to ukázat, jak kvalitní infrastruktura, která na děti myslí, může podpořit jejich potenciální růst, a pokud se k tomu přidá i zajímavá finanční podpora, která umožní matce být s dítětem na mateřské první roky jeho života, které jsou klíčové pro jeho rozvoj, může to mít pozitivní vliv.

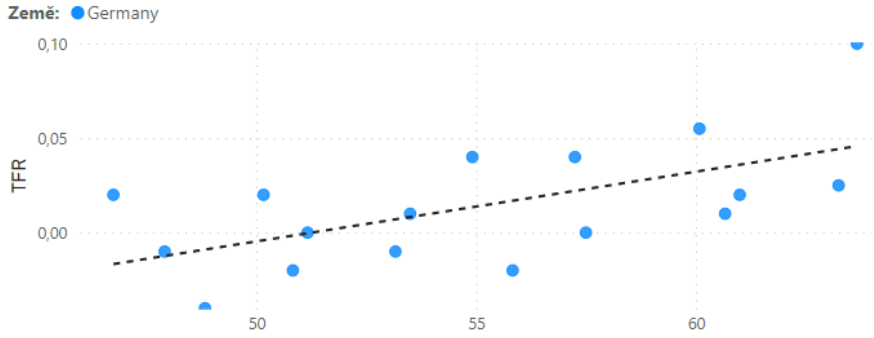
6.4.4 Psi a kočky

Poslední skupinou, které bych se chtěl blíže věnovat, jsou domácí mazlíčci, lépe řečeno psi a kočky. U této skupiny jsem se domníval, že budou mít spíše negativní efekt,neboť je spousta lidí, kteří si právě těmito mazlíčky děti vynahrazují, a to pak na samotnou porodnost dopadá.

Ovšem problém této skupiny je mnohem složitější, jak jsem při jejím bližším zkoumáním zjistil, a dost se liší stát od státu. Hlavní příčinou může být kulturní odlišnost a to, jak dané země tato zvířata vnímají, ale hlavně i samotný vztah lidí k daným zvířatům a představě rodiny všeobecně. Pomineme-li ovšem ty země, kde nejsou tato zvířata brána jako domácí mazlíčci, a tedy nemá smysl jejich vliv na korelaci v těchto zemích zkoumat, tak se zdá, že se lidé dělí na dva tábory.

První skupina přesně zapadá do mého předpokladu a zde je vliv dle očekávání negativní a i z výsledků se zdá, že průměrně mazlíčci počet dětí snižují, ale pak je druhá skupina lidí, kteří naopak preferují klasickou představu rodiny, kde k dětem patří právě i domácí mazlíček, jenž je plnohodnotným členem domácnosti.

Díky této rozpolcenosti jsou pak krásně vidět rozdíly, kdy například v USA má pes poměrně vysokou negativní korelaci, ale naopak v Německu je silná pozitivní korelace. A díky tomu je i spousta zemí, kde jsou oba tábory dohromady, a díky tomu jsou následně výsledky statisticky neprůkazné a z hodnocení vypadají stejně jako země, kde tato zvířata jako mazlíčky neřeší. Proto je zrovna v tomto případě myslet na tyto další faktory, ale samo o sobě asi nelze brát domácí zvířata jako značné hybatele TFR a jako něco, díky čemu by se mohla její hodnota zvýšit.

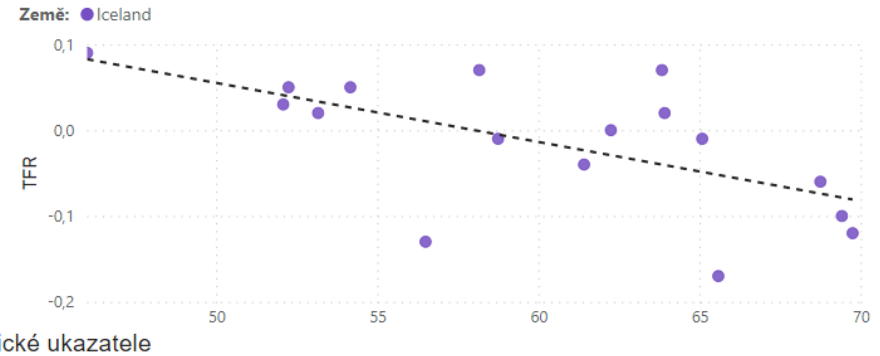


Statistické ukazatele

Y = a + bX + e

X	Y	a	b	e	p	R	R2
TFR	Value	53.69724	98.52902	33.66104	0.01041	0.60294	0.36354

Obrázek 6.3: Ukázka vlivu psů na TFR v Německu, kde má významný pozitivní efekt na TFR.



Statistické ukazatele

Y = a + bX + e

X	Y	a	b	e	p	R	R2
TFR	Value	59.29004	-54.80262	18.18492	0.00873	-0.61411	0.37713

Obrázek 6.4: Ukázka vlivu psů na TFR na Islandu, kde má významný negativní efekt na TFR.

Závěr

V této práci jsem se zabýval analýzou vztahu mezi TFR a mnou vybranými veličinami za pomocí metodiky BI. Z tohoto hlediska bylo nejprve potřeba čtenáři osvětlit používané pojmy a celkovou problematiku TFR. Dále také matematické metody, které byly využity při statistickém vyhodnocování a v poslední řadě byla čtenáři představena metodika BI a veškeré náležitosti, jež s ní souvisí a byly využity pro tvorbu této práce.

V praktické části byl navržen a vytvořen datový sklad, ktarý za pomoci automatických skriptů a ETL procesů zpracovává zvolené datové sady do požadovaného formátu (dimenzionální databáze). Z této databáze jsem následně v nástroji Power BI vytvořil interaktivní dashboard, v kterém je možné za pomocí přednastavených filtrů libovolně zkoumat vztahy zvolených veličin. Tento nástroj jsem následně publikoval na webové stránky, které jsem za tímto účelem vytvořil, abych tím umožnil přístup co nejvíce uživatelům a mohl tak šířit větší povědomí o této problematice.

Na závěr jsem dané veličiny sám statisticky vyhodnotil a zároveň jsem uvedl zajímavé poznatky, které z bližší analýzy ve vytvořeném dashboardu vyplynuly.

Součástí práce jsou také veškeré zdrojové kódy a manuál, který je určen pro uživatele, kteří by si rádi tento datový sklad postavili u sebe na stanici. Také je zde návod, jak přidat další datové sady, které dané uživatelé zajímají, a zkoumat tak nové možnosti.

Literatura

[1] *Determinants and Consequences of High Fertility: A Synopsis of the Evidence*, [online], 2010, [cit. 2022-09-17], Dostupné z: https://openknowledge.worldbank.org/bitstream/handle/10986/27497/630690WP0P10870nants0pub08023010web.pdf?fbclid=IwAR2mLj5BF1I13tUA7SrMKAQOJKwKBCwU7toPE5j6fkoktahHV5nmraGeNfA

[2] *United Nations, Population Division, United Nations publication, Sales No. E.06.XIII.5, New York, 2006*, [cit. 2022-09-17], Dostupné z: https://www.un.org/esa/sustdev/natlinfo/indicators/methodology_sheets/demographics/total_fertility_rate.pdf

[3] *Götmark, F., Andersson, M. Human fertility in relation to education, economy, religion, contraception, and family planning programs.* BMC Public Health 20, 265, 2020. https://doi.org/10.1186/s12889-020-8331-7

[4] *CHOI, Yoonjoung, Madeleine SHORT FABIC a Jacob ADETUNJI. Does age-adjusted measurement of contraceptive use better explain the relationship between fertility and contraception?*. Demographic Research [online]. 2018, 39, 1227-1240 [cit. 2022-09-25]. ISSN 1435-9871. Dostupné z: doi:10.4054/DemRes.2018.39.45

[5] Girum *T, Wasie A. Return of fertility after discontinuation of contraception: a systematic review and meta-analysis.* Contracept Reprod Med. 2018 Jul 23,3:9. [cit. 2022-09-21], doi: 10.1186/s40834-018-0064-y. PMID: 3006204, PMCID: PMC6055351.

[6] *Yujie, Li, The Relationship between Fertility Rate and Economic Growth in Developing Countries*, Supervisor: Martin Dribe, [Master thesis], 2015, [cit. 2022-09-25], Dostupné z: https://lup.lub.lu.se/luur/download?func=downloadFile&recordOId=8727479&fileOId=8768892&fbclid=IwAR1e53pDkNj2nRrq_tmOOiIDv2IQp2q33yziVQD4iaIsJWPNV0ADcNfklo

[7] *Pourreza, A., Sadeghi, A., Amini-Rarani, M. et al. Contributing factors to the total fertility rate declining trend in the Middle East and North Africa: a systemic review.* J Health Popul Nutr 40, 11 (2021).[cit. 2022-09-25], https://doi.org/10.1186/s41043-021-00239-w

[8] *EVAN, Tomáš a Pavla VOZÁROVÁ. Influence of women's workforce participation and pensions on total fertility rate: a theoretical and econometric study.* Eurasian Economic Review [online]. 2018, 8(1), 51-72 [cit. 2022-09-25]. ISSN 1309-422X. Dostupné z: doi:10.1007/s40822-017-0074-0

[9] *Martine, George, Jose Eustaquio Alves, and Suzana Cavenaghi. Urbanization and Fertility Decline: Cashing in on Structural Change.* International Institute for Environment and Development, 2013. [cit. 2022-09-24], http://www.jstor.org/stable/resrep01293.

[10] *BROTZ, Daniel. Hledání a práce s veličinami souvisejícími s TFR.* Praha, 2022. Bakalářská práce. ČVUT Fakulta informačních technologií. Vedoucí práce Evan Tomáš.

[11] *Kuznets, Simon. "Rural-Urban Differences in Fertility: An International Comparison."* Proceedings of the American Philosophical Society, vol. 118, no. 1, 1974, pp. 1–29. JSTOR, [cit. 2022-09-25], http://www.jstor.org/stable/986434.

[12] *United Nations, Population Division, FERTILITY LEVELS AND TRENDS IN COUNTRIES WITH INTERMEDIATE LEVELS OF FERTILITY*, 2002, [cit. 2022-09-17], Dostupné z: https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/unpd_egm_200203_backgroundpaper_fertility_levels_and_trends_population_division.pdf?fbclid=IwAR1KiGoqZgvyzqFxhuB_QcIbZVPZQ9uGmMNHcMCDDESWKzksB9krimSA5Ew

[13] *KŘIVÝ, Ivan. ANALÝZA ČASOVÝCH ŘAD* [online]. Ostrava, 2012 [cit. 2022-09-29]. Dostupné z: https://web.osu.cz/~Bujok/files/ancas.pdf. Skripta. Ostravská univerzita v Ostravě.

[14] *BLAŽEK, B., Jitka HRABÁKOVÁ, Pavel HRABÁK, Roman KOTECKÝ, Petr NOVÁK a Daniel VAŠATA. Testování hypotéz: 10. přednáška* [online]. Praha, 2011 [cit. 2022-09-29]. Dostupné z: https://courses.fit.cvut.cz/NI-VSM/lectures/files/NI-VSM-Lec-10-Slides.pdf. Učební materiál. ČVUT.

[15] *BLAŽEK, B., Jitka HRABÁKOVÁ, Pavel HRABÁK, Roman KOTECKÝ, Petr NOVÁK a Daniel VAŠATA. Náhodné vektory: 3. přednáška* [online]. Praha, 2011 [cit. 2022-09-29]. Dostupné z: https://courses.fit.cvut.cz/NI-VSM/lectures/files/NI-VSM-Lec-03-Slides.pdf. Učební materiál. ČVUT.

[16] *Jednoduchá lineární regrese* [online]. Brno, 2022 [cit. 2022-09-29]. Dostupné z: https://mathstat.econ.muni.cz/media/19031/linearni-regrese.pdf. Učební materiál. Masarykova univerzita.

[17] *Základy ekonometrie: IX. Analýza jednorozměrných časových řad* [online]. Brno, 2015 [cit. 2022-09-29]. Dostupné z: https://is.muni.cz/el/1456/podzim2015/BKE_ZAEK/um/59125597/09_CasRady.pdf. Učební materiál. Masarykova univerzita.

[18] *POKORNÝ, Martin. Testy jednotkového kořene a jejich využití v ekonomii* [online]. Brno, 2012 [cit. 2022-09-29]. Dostupné z: https://is.muni.cz/th/uzm4g/DIPLOMOVA_PRACE.pdf. Diplomová práce. Masarykova univerzita. Vedoucí práce Ing. Daniel Němec, Ph.D.

[19] *KOLÁŘ, Robert a Jakub KREJČÍ. Podnikové datové sklady: 1. přednáška* [online]. Praha, 2022 [cit. 2022-09-29]. Dostupné z: https://courses.fit.cvut.cz/NI-EDW/lectures/01_prednaska.pdf. Učební materiál. ČVUT.

[20] *KOLÁŘ, Robert a Jakub KREJČÍ. Podnikové datové sklady: 2. přednáška* [online]. Praha, 2022 [cit. 2022-09-29]. Dostupné z: https://courses.fit.cvut.cz/NI-EDW/lectures/02_prednaska.pdf. Učební materiál. ČVUT.

[21] *KOLÁŘ, Robert a Jakub KREJČÍ. Podnikové datové sklady: 1. cvičení* [online]. Praha, 2022 [cit. 2022-09-29]. Dostupné z: https://courses.fit.cvut.cz/NI-EDW/tutorials/01_cviceni.pdf. Učební materiál. ČVUT.

[22] *KOLÁŘ, Robert a Jakub KREJČÍ. Podnikové datové sklady: 3. přednáška* [online]. Praha, 2022 [cit. 2022-09-29]. Dostupné z: https://courses.fit.cvut.cz/NI-EDW/lectures/03_prednaska.pdf. Učební materiál. ČVUT.

[23] *KOLÁŘ, Robert a Jakub KREJČÍ. Podnikové datové sklady: 4. přednáška* [online]. Praha, 2022 [cit. 2022-09-29]. Dostupné z: https://courses.fit.cvut.cz/NI-EDW/lectures/04_prednaska.pdf. Učební materiál. ČVUT.

[24] *KOLÁŘ, Robert a Jakub KREJČÍ. Podnikové datové sklady: 5. přednáška* [online]. Praha, 2022 [cit. 2022-09-29]. Dostupné z: https://courses.fit.cvut.cz/NI-EDW/lectures/05_prednaska.pdf. Učební materiál. ČVUT.

[25] *KAPUSTOVÁ, Veronika. Cvičební pomůcka pro předmět Základy statistiky pro přírodní vědy.* Ostrava 2021, [cit. 2022-09-29], Učební materiál. Ostravská univerzita v Ostravě.

[26] *Download Python | Python.org. Welcome to Python.org* [online]. Copyright ©2001 [cit. 23.8.2022]. Dostupné z: https://www.python.org/downloads/

[27] *PostgreSQL: Downloads. PostgreSQL: The world's most advanced open source database* [online]. Copyright © 1996 [cit. 23.8.2022]. Dostupné z: https://www.postgresql.org/download/

[28] *Postbird | Apps | Electron. Electron | Build cross-platform desktop apps with JavaScript, HTML, and CSS.* [online]. Dostupné z: https://electron-website.herokuapp.com/apps/postbird

[29] *Full Lifecycle Modeling for Business, Software and Systems | Sparx Systems.* UML modeling tools for Business, Software, Systems and Architecture [online]. Copyright © 2000 [cit. 28.8.2022]. Dostupné z: https://sparxsystems.com/products/ea/

[30] *Installing Pentaho Data Integration CE | Hitachi Vantara.* [online]. Copyright © Hitachi Vantara LLC 2022. All Rights Reserved. [cit. 28.8.2022]. Dostupné z: https://www.hitachivantara.com/en-us/pdf/implementation-guide/three-steps-to-install-pentaho-data-integration-ce.pdf

[31] *Visual Studio Code - Code Editing. Redefined. Visual Studio Code - Code Editing. Redefined* [online]. Copyright © 2022 Microsoft [cit. 23.8.2022]. Dostupné z: https://code.visualstudio.com/

[32] textitVizualizace dat | Microsoft Power BI. Object moved [online]. Copyright © 2022 Microsoft [cit. 31.8.2022]. Dostupné z: https://powerbi.microsoft.com/cs-cz/

[33] *pytrends · PyPI. PyPI · The Python Package Index* [online]. Copyright © 2022 [cit. 28.8.2022]. Dostupné z: https://pypi.org/project/pytrends/

[34] *country-converter · PyPI. PyPI · The Python Package Index* [online]. Copyright © 2022 [cit. 28.8.2022]. Dostupné z: https://pypi.org/project/country-converter/

[35] *Introduction — statsmodels.* [online]. [cit. 28.8.2022]. Dostupné z: https://www.statsmodels.org/stable/index.html

[36] *NumPy. NumPy* [online]. Copyright © 2022 NumPy. All rights reserved. [cit. 28.8.2022]. Dostupné z: https://numpy.org/

[37] *pandas - Python Data Analysis Library. pandas - Python Data Analysis Library* [online]. Copyright © 2022 pandas via [cit. 28.8.2022]. Dostupné z: https://pandas.pydata.org/

[38] *Google Trends.* [online].[cit. 24.7.2022]. Dostupné z: https://trends.google.com/trends/?geo=CZ

[39] *World Bank Open Data | Data. World Bank Open Data | Data* [online]. Copyright © [cit. 24.7.2022]. Dostupné z: https://data.worldbank.org/

[40] *Our World in Data. Our World in Data* [online]. [cit. 28.7.2022]. Dostupné z: https://ourworldindata.org/

[41] *Mean body mass index (BMI) in women - Our World in Data. Our World in Data* [online]. [cit. 29.7.2022]. Dostupné z: https://ourworldindata.org/grapher/mean-body-mass-index-bmi-in-adult-women

[42] *P90 vs. P10 of income/consumption distribution - Our World in Data. Our World in Data* [online]. [cit. 29.7.2022]. Dostupné z: https://ourworldindata.org/grapher/p90-vs-p10-logs

[43] *Average years of schooling for women - Our World in Data. Our World in Data* [online]. [cit. 29.7.2022]. Dostupné z: https://ourworldindata.org/grapher/mean-years-of-schooling-female

[44] *Number of people living in urban and rural areas - Our World in Data. Our World in Data* [online]. [cit. 29.7.2022]. Dostupné z: https://ourworldindata.org/grapher/urban-and-rural-population

[45] *Annual working hours per worker - Our World in Data. Our World in Data* [online]. [cit. 29.7.2022]. Dostupné z: https://ourworldindata.org/grapher/annual-working-hours-per-worker

[46] *Death rate from natural disasters - Our World in Data. Our World in Data* [online]. [cit. 29.7.2022]. Dostupné z: https://ourworldindata.org/grapher/death-rates-from-disasters

[47] *Fertility rate, total (births per woman) | Data. World Bank Open Data* | Data [online]. Copyright © [cit. 26.7.2022]. Dostupné z: https://data.worldbank.org/indicator/SP.DYN.TFRT.IN

[48] *Labor force, female (% of total labor force) | Data. World Bank Open Data* | Data [online]. Copyright © [cit. 26.7.2022]. Dostupné z: https://data.worldbank.org/indicator/SL.TLF.TOTL.FE.ZS

[49] *World Fertility Data | Population Division. Welcome to the United Nations* [online]. Dostupné z: https://www.un.org/development/desa/pd/data/world-fertility-data

Příloha A

Seznam použitých zkratek

API Application Programming Interface

AR Autoregresivní

BI Business intelligence

BMI Body Mass Index

CSV Comma-separated values

DDL Data Definition Language

DI Data integration

DWH Data Warehouse

ETL Extract, Transform, Load

IDE Integrated Development Environment

iid Independent and identically distributed

IP Internet protocol

KPI Key performance indikator

KPSS Kwiatkowski–Phillips–Schmidt–Shin

LM Lagrangeův multiplikátor

OLS Ordinary least squares

SQL Structured Query Language

TFR Total fertility rate

Příloha B

Malual

Tato příloha má uživateli pomoci správně nastavit své prostředí a navést ho při práci s ETL nástrojem Pentaho Data Integration.

B.1 Vytvoření DWH

Tato část obsahuje podrobný návod pro vytvoření celého datového skladu v takovém stavu, v jakém je zhotoven pro potřeby této práce.

B.1.1 Softwarové požadovky

Aby bylo možné vytvořit celý sklad automaticky pomocí Pentaho Jobu Create_DWH, je třeba mít na zařízení následující nástroje:

- PostgreSQL databáze
  - Stage
  - Target
- Python 3.0+ [26](+knihovny)
  - Pandas
  - country-converter
  - pytrends
  - numpy
  - statsmodels
- Pentaho Data Integration 6.0.0.0-353+
- Power BI

B.1.2 Nastavení proměnné prostředí

V nástroji pro správu ETL procesů je třeba nastavit správně proměnné prostředí, které slouží pro správnou práci se soubory a skripty. Postup je následující: Zvolte **Edit -> Set Enviromental Variables**. Zde je možné nastavit dvě proměnné:

1. **path** - kam vložíte celou cestu k odresáři, kde je vložena kompletní složka DP-data (např. C:\Program Files (x86))
2. **python** - zde vložíte cestu k python kompilatoru, který má přístup ke všem potřebným knihovnám

   (např. C:\Users\AppData\Local\Microsoft\WindowsApps\python.exe)

B.1.3 Připojení databáze

Pro správnou funkčnost je třeba změnit připojení obou dílčích databází na ty, které máte vytvořeny na vlastním stroji. Postup je následující:

1. Otevřete Pentaho Job Create_DWH
2. Na boční Tool liště zvolte **View->Jobs->Create_DWH->Database connections**
3. Zde klikněte pravým tlačítkem na jednotlivé FR připojení, zvolte možnost **edit** a následně v okénku wizzardu nastavte připojení na vlastní databazi Stage či Target

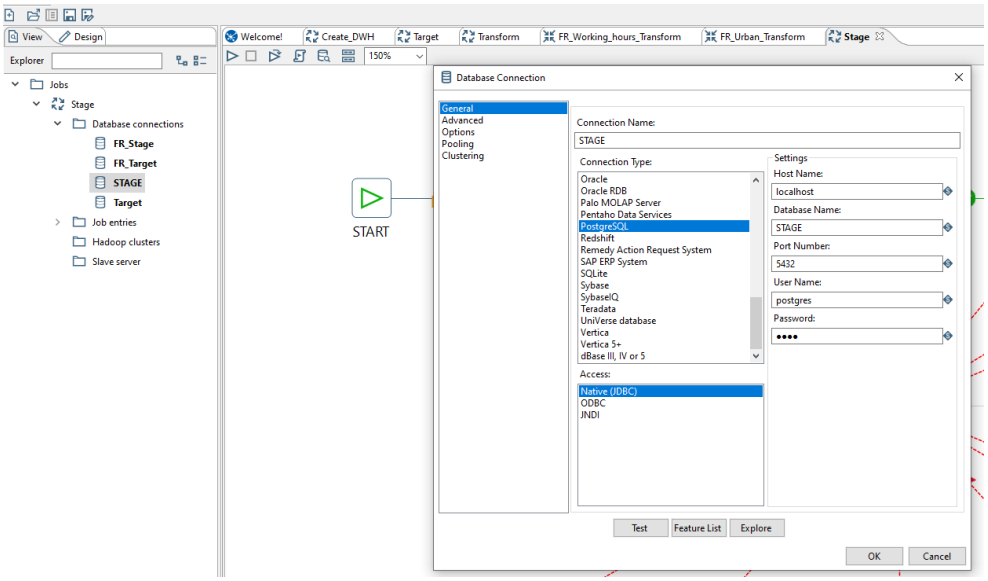
B.1.4 Propojení Power BI s databází

Po otevření již připraveného FR_Dashboardu by měly být data součástí sestavy. Pokud chcete data aktualizovat nebo dashboard napojit přímo na vlastní databázi, stačí postupovat dle popisu dále:

1. V záložce **Domů\Home -> Získat data\Get data -> Více\More -> Databáze\Database -> databáze PostgreSQL**
2. Dále vyplňte údaje potřebné pro připojení vlastní databáze a zvolte potřebné tabulky náhledů, ze kterých je vytvořen datamart a z jehož dat se vytváří vlastní vizualizace.

B.2 Přidání datové sady do DWH

Datový soubor by měl mít příponu .csv, a být uložen ve složce:
DP-data\Original_Data_Sources.



Obrázek B.1: Obrázek k návodu k připojení databází v Pentaho DI.

Pentaho DI samozřejmě podporuje i jiné typy datových souborů a lze je tedy bez problému nahrát, ale v tomto případě nelze využít možnosti se v nastavení jednotlivých částí transakce inspirovat již předpřipravenými částmi.

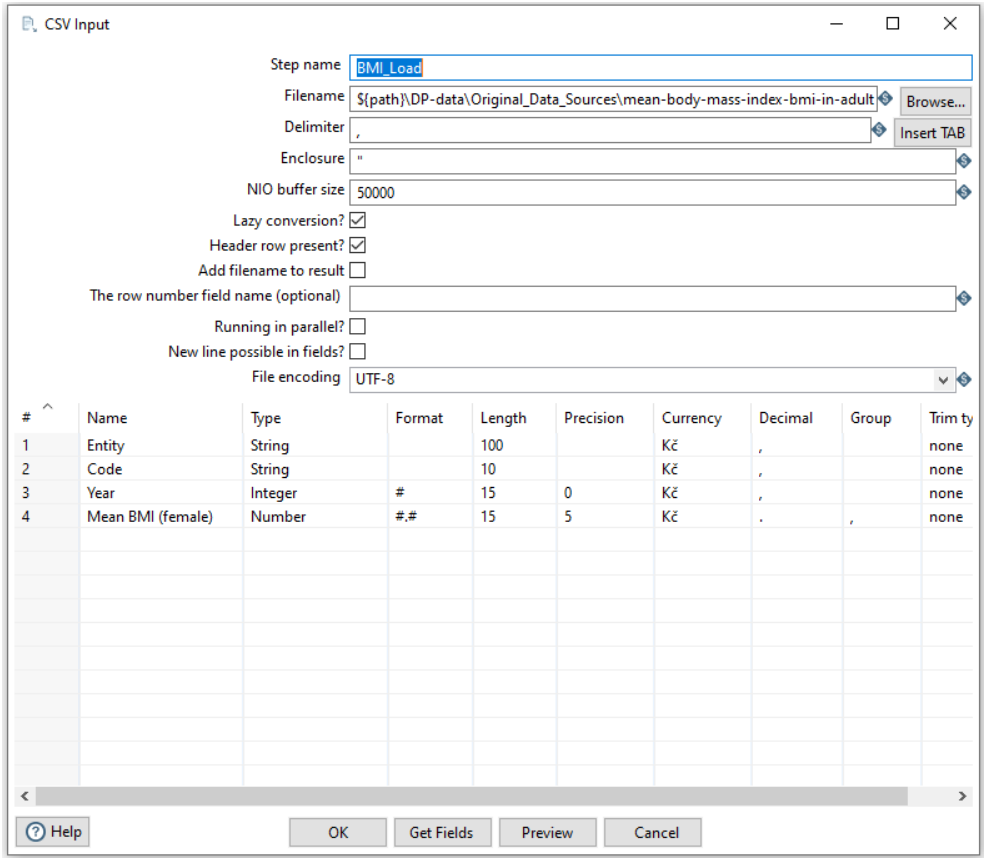
Do SQL souboru FR_Stage lze doplnit část pro vytvoření tabulky pro daný soubor a nebo je možné nechat vytvořit tabulku přímo z Pentaho DI, který rovnou i zobrazí kompletní SQL kód, který lze následně do tohoto souboru jen zkopírovat pro případné znovuvytvoření celého skladu.

Zbylé tabulky již nejsou při přidávání souboru ovlivněny, a tak stačí již pouze správně vytvořit ETL transakce a přiřadit je do Správného Jobu.

B.2.1 Vytvoření transakce pro Stage

Po zvolení možnosti **Vytvořit novou transakci** je nutné z toolboxu zvolit **CSV file input**, v něm vybrat vámi stažený soubor. Popřípadě zvolit jinou možnost inputu dat, dle vaší situace. Po vložení správného souboru a zvolení správného nastavení dle vlastností dat a souboru, lze vybrat **Get Fields**. Nahrají se všechny sloupce dle hlaviček CSV souboru. Následně je třeba upravit datové typy a nejspíš bude třeba u určitých hodnot zvolit vyšší hodnotu **Length**, aby všechny data bez problému prošla, protože přednastavené velikosti jsou určeny pouze z menšího vzorku a ne celého souboru, a tak nemusí odpovídat realitě.

Následně zvolte z Toolbaru **Table output** a spojte s předchozím krokem ve správném směru šipky. Připojte k ní **Stage** databázi s tabulkou pro tyto data. Zvolte **Truncate table** a **Specify database fields**. Následně klikněte



Obrázek B.2: Ukázka nastavení hodnot pro CSV input.

na **Database fields**, kde zvolíte **Get fields**.

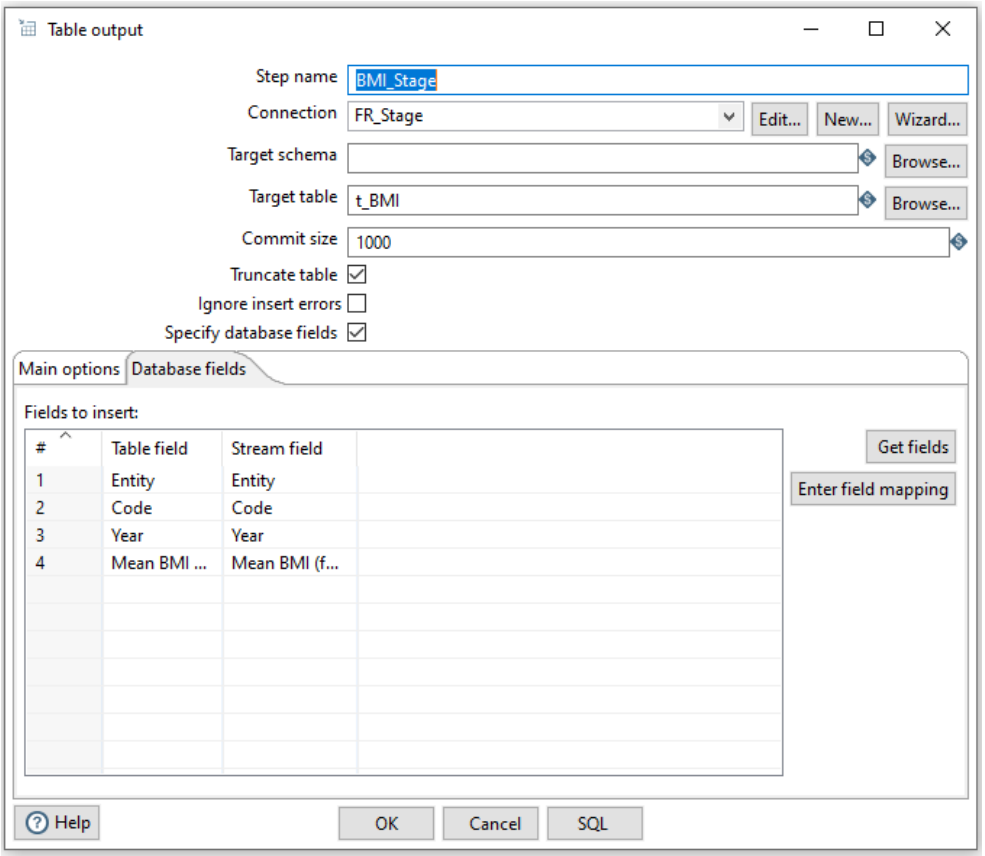
Pokud nemáte již připravenou tabulku, stačí zvolit možnost **SQL** a vygeneruje se kód pro její vytvoření a zároveň je tabulka s vámi zvoleným názvem (ten by neměl kolidovat s již vzniklými tabulkami, jinak budou přepsány) vytvořena na databázovém stroji **Stage**.

Následně stačí přidat tuto transakci do Stage jobu, aby bylo možné vše automatizovat. (Při vytváření je ideální si otevřít jinou stage transakci a nechat se jí inspirovat).

B.2.2 Vytvoření transakce pro Transform

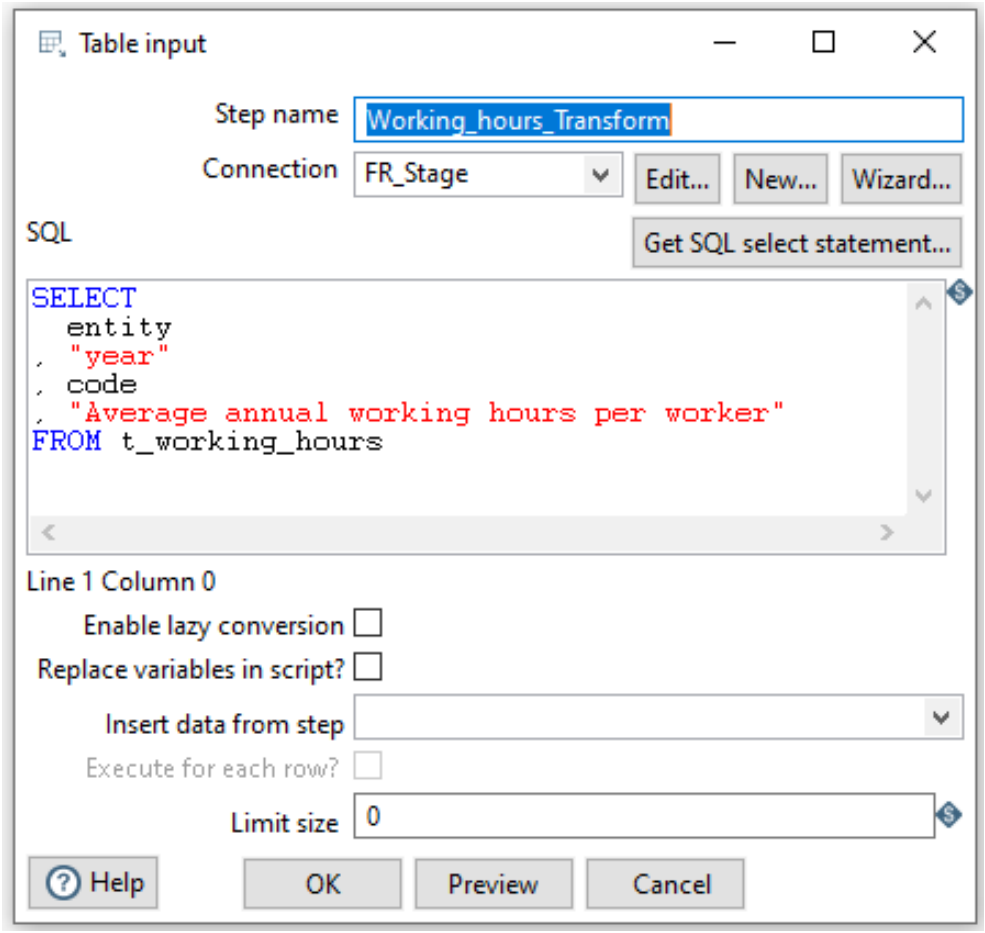
Nejdříve vytvoříme novou transakci a do ní vložíme **Table input**, ke kterému připojíme **Stage** databázi. Zvolíme možnost **Get SQL select statement...** Ve výběru zvolíme tabulku, z které chceme vzít data a vezmeme z ní všechny sloupce.

Další úpravy jsou závislé na tom, jak jsou všechny data v dané datové sestavě uložena, ale vždy je třeba zvolit **Select values**, kde všechny nepo-



Obrázek B.3: Ukázka nastavení v Table output.

Obrázek B.4: Ukázka šablony prvků pro Stage transakci.



Obrázek B.5: Ukázka nastavení v Table input.

třebné sloupce zařadíme do kolonky **Remove** a zbylé sloupce přejmenujeme dle následné šablony:

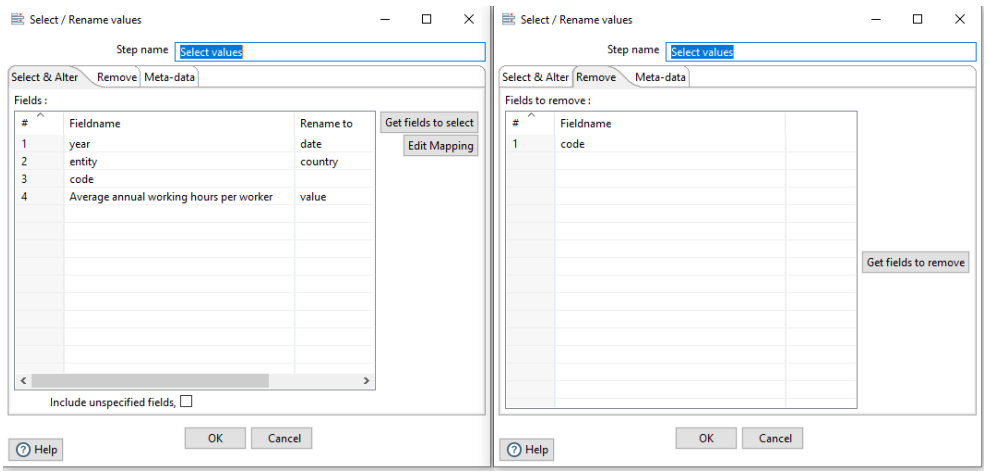
country - sloupec pro název země

date - sloupec pro časový údaj ve tvaru yyyy a formátu int

value - sloupec pro vlastní hodnotu měřené veličiny

Poslední nutnou úpravou je uložená tohoto datového souboru do CSV do složky **transform**, z níž budou všechny CSV soubory transformovány pomocí Python skriptu **data_transform.py**, který upraví názvy zemí, aby byly jednotné, a otestuje stacionaritu dat.

K tomu je potřeba přidat **Text file output**. V něm zvolit na úvodní stránce správnou cestu, kam se má soubor uložit ($path\DP-data\transform\+ náš název souboru bez přípony). V **Extension** zvolíte csv.



Obrázek B.6: Ukázka nastavení v Select values a Remove.

Obrázek B.7: Ukázka šablony prvků pro Transform transakci.

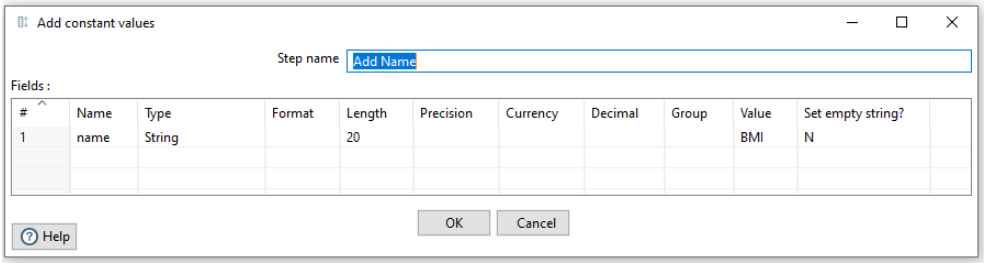
V části **Content** zvolíte jako Separator ; a jako Enclosure “. Potvrdíte Header a pokračujete na kolonku **Fields**, kde opět přes **Get Fields** dostanete obsah a následně zvolíte možnost **Minimal width**, aby se nuceně neroztahovaly stringy.

Výsledek je následně uložen do souboru **Converted_Data_Sources**, z něhož se následně načítají data v dalším kroku. Po dokončení je třeba tuto transformaci přidat do **Transform** Pentaho Jobu, aby byl tento krok zařazen do automatizovaného procesu tvorby datového skladu.

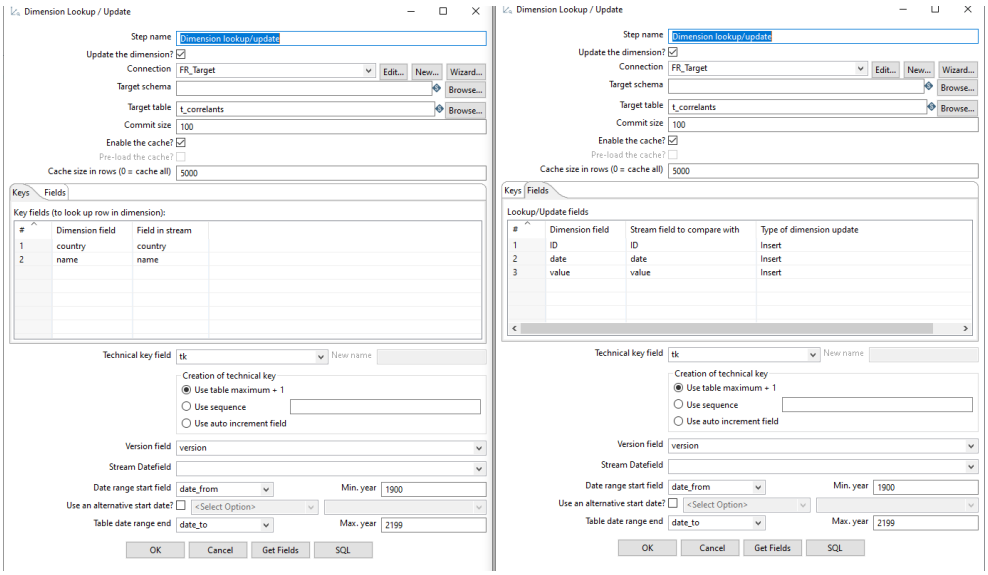
B.2.3 Vytvoření transakce pro Target

Opět je třeba nejdříve vytvořit transakci, v níž zvolíte **CSV file input**, kde vybereme náš transformovaný soubor z **Converted_Data_Sources**. Pokud tento soubor ještě nemáme transformovaný, pouze vytvoříme cestu, kde se bude nacházet a samotné sloupce doplníme dle šablony z jiné Target transakce, neboť v této fázi mají již všechny soubory jednotný tvar. Nastavení této položky je stejné, jako v případě části pro Stage. Jen je potřeba do první kolonky vložit název **ID**ID, protože ten není obsažen v headeru CSV souboru.

Dále je potřeba přidat sloupec **Name** s názvem proměnné, což lze provést možností **Add constants**, v níž zvolíme **Type string** a **Value** dle názvu



Obrázek B.8: Ukázka nastavení v add constants pro přidaní sloupce Name.



Obrázek B.9: Ukázka nastavení Dimension lookup pro přidání dat do Target databáze a jejich historizaci.

naší proměnné. I když by to nemělo být již třeba, lze provést další nezbytné úpravy, aby následná data odpovídala tabulce **t_correlants** bez pomocných sloupců.

Následně přidejte z toolboxu **Dimension lookup/update**, kde je opět třeba připojit nyní **Target** databázi. A doplnit správně všechny parametry technických proměnných. **Key fields** by měli být vždy sloupce pro Zemi a Název proměnné (Country, Name). Type of dimensional update je **Insert** a názvy sloupců by měly odpovídat názvům v **Target** databázi a není třeba již nic upravovat.

Následně je opět třeba přidat danou transakci do Target Jobu (FR_Target) a je velmi vhodné postupovat dle jiné target transakce na datech podobných.

Obrázek B.10: Ukázka šablony prvků pro Target transakci.

Příloha C

Obsah přiloženého CD

- `readme.txt` ..... stručný popis obsahu
- `FR_Dashboard` ..... Power BI dashboard s data použitá v této práci
- `DP-data`
  - `Original_Data_Sources` ..... originální datové sady
  - `Google_Trans` ..... dílčí části datových sad stažených z Google trends
  - `transform` ..... pomocné soubory sloužící pro následnou transformaci
  - `Converted_Data_Sources` ..... transformovaná data
  - `Pictures` ..... obrázky modelů, grafů a teoretických ukázek
  - `Python_scripts` ..... python skripty
  - `SQL` ..... SQL soubory pro vytvoření datového skladu
  - `Pentaho_Transactions` ..... zdrojové kódy pro pentaho transakce
  - `Pentaho_Jobs` ..... zdrojové kódy pro pentaho joby
  - `DP-TFR` ..... zdrojový kód pro Enterprase Architect
- `thesis`
  - `thesis_latex` ..... zdrojová data práce ve formatu LaTeX
  - `thesis.pdf` ..... text práce ve formátu PDF