**Výzkumný ústav bezpečnosti práce, v. v. i.**
JERUZALÉMSKÁ 1283/9
110 00 PRAHA 1 – NOVÉ MĚSTO

# Podpůrná metodika k aktualizaci webové aplikace „Statistiky BOZP“

Číslo výzkumného úkolu: VUS4_03_VÚBP

Název výzkumného úkolu:
**NOVÉ PŘÍSTUPY PRO TVORBU STATISTIK A ZLEPŠOVÁNÍ INFORMAČNÍ ZÁKLADNY O BOZP**

Hlavní řešitel: Výzkumný ústav bezpečnosti práce, v. v. i.

**Výzkumný ústav bezpečnosti práce, v. v. i.**
JERUZALÉMSKÁ 1283/9
110 00 PRAHA 1 – NOVÉ MĚSTO

# Datové zdroje

# 1. Úvod

Úkolem projektu bylo vytvořit datový portál s popisnou analýzou údajů spadající pod činnost ***BOZP***. Portál ze své podstaty musí vycházet z dat relevantních pro danou oblast.

Ve stávající podobě jsou analýzy založeny na datech ze čtyř oblastí:

– pracovní úrazy,
– invalidita,
– dočasná pracovní neschopnost (agregované hodnoty),
– pracovní neschopnost (individuální případy).

Zdrojová data byla získána od různých poskytovatelů v rozličných formátech a kvalitě, přehled dat je uveden v Tab. 1. Data byla doplněna veřejně dostupnými statistickými údaji.

Aby data bylo možné použít jako kvalitní podklad pro portál, musela projít rozsáhlou úpravou. Popis zdrojových dat a jejich úpravy je předmětem tohoto materiálu. Oblasti dat mají různou složitost, což se odrazilo v mírně odlišné struktuře jednotlivých kapitol.

Pro datové úpravy byly použity nástroje ***MS Excel***, ***IBM SPSS Modeler*** a ***IBM SPSS Statistics***.

*Tab. 1 Datové zdroje*

| Poskytovatel dat | Název datového souboru | Období |
|---|---|---|
| Státní úřad inspekce práce, Český báňský úřad | Pracovní úrazy | 2014–2019 |
| Česká správa sociálního zabezpečení | Dočasná pracovní neschopnost | 2014–2018 |
| Česká správa sociálního zabezpečení | Případy invalidity | 2014–2018 |
| Ústav zdravotnických informací a statistiky | Registr pracovní neschopnosti | 2014–2017 |
| Český statistický úřad | Počty obyvatel a pracovníků | 2014–2018 |

# 2. Pracovní úrazy

Data o pracovních úrazech pocházejí z období 2014–2019, jejich poskytovatelem je **Státní úřad inspekce práce**. Data byla rozšířena pro roky 2017–2019 o údaje z **Českého báňského úřadu**. Vstupní data jsou v nativním formátu ***SSPS Statistics*** *(.sav)* s výjimkou let 2015 a 2016, kde jsou data ve formátu ***MS Excel*** *(.xls)*. Řádek v datové matici představuje jeden pracovní úraz s dobou pracovní neschopnosti delší než 3 dny.

## 2.1. Proměnné

Struktura souborů se v jednotlivých letech lišila, některé proměnné se vyskytovaly jen v některých letech. Měnila se také jména proměnných. Po dohodě s poskytovatelem dat byly využity následující proměnné, které jsou pro snazší orientaci rozděleny do tří oblastí:

– Identifikace úrazu
  - Zdroj dat – instituce poskytující data
  - Datum úrazu – datum, kdy došlo k úrazu <1.1.2014-31.12.219>
  - okres úrazu – okres zaměstnavatele

– Informace o osobě s úrazem
  - Věk – věk v letech
  - Pohlaví - pohlaví
  - CZNACE – odvětví zaměstnavatele dle číselníku
  - KZAM – profese zaměstnance dle číselníku, data do roku 2018
  - ISCO – profese zaměstnance dle číselníku, data od roku 2019
– Specifikace úrazu
  - Druh úrazu – závažný úraz a ostatní úraz, dle číselníku
  - Druh zranění – druh zranění dle číselníku
  - Část těla – zraněná část těla dle číselníku
  - Místo úrazu – určení místa, kde došlo k úrazu, dle číselníku
  - Činnost při úrazu – činnost vykonávaná v době úrazu, dle číselníku
  - Zdroj úrazu – základní materiální činitel související s kontaktem při zranění, objekt, nástroj, nebo zařízení, se kterým přišel pracovník do styku, hodnoty dle číselníku

## 2.2. Datové operace

Výsledný datový zdroj obsahuje spojená data za celé období. Protože spojení souborů je podmíněno stejnou strukturou souborů s jednotným formátem a významem proměnných, bylo nutné nejprve strukturu sjednotit. Po sjednocení následovaly opravy hodnot, doplnění dalších proměnných a připojení hodnot z číselníků. Úpravy byly dokončeny uspořádáním proměnných doplněním metadat a uložením do formátu *.sav* (***SPSS Statistics***). Všechny datové operace probíhaly v nástroji ***SPSS Modeler***.

### 2.2.1. Sjednocení struktury a spojení

Při sjednocování struktury bylo nutné nejprve zvolit proměnné se stejným významem, jednotně je přejmenovat a sjednotit jejich formáty.

Proměnné nesoucí podobnou informaci byly v souborech v různém tvaru. Například proměnné ze skupiny týkající se okresu v různých souborech zachycovaly sídlo zaměstnavatele, místo bydliště zraněné osoby nebo místo, kde došlo k pracovnímu úrazu. Proměnné se stejným významem se v různých letech nazývaly různě. V některých souborech byl uveden kód okresu, jindy název. Výsledné proměnné byly vybrány na základě konzultací s poskytovatelem dat. Proměnné byly jednotně pojmenovány a byl změněn jejich datový formát (např. formát celého čísla, nebo datum).

Do dalšího zpracování postoupily proměnné, které se vykytovaly ve všech souborech. Pro analýzu jsou dobře použitelné pouze proměnné, které pokrývají celé sledované období. Výjimku tvoří proměnné týkající se profese *KZAM* a *ISCO*, kde došlo ke změně používaného číselníku. Proměnná *KZAM* je vyplněna do roku 2018 a proměnná *ISCO* od roku 2019. V ostatních letech se v datech vyskytují chybějící hodnoty.

Po sjednocení a výběru proměnných došlo ke spojení soborů. Napojeny na sebe byly proměnné se stejnými názvy.

### 2.2.2. opravy chybných hodnot a doplnění

Protože většina proměnných byla textová, obsahovaly opravy zejména nahrazování překlepů a odstraňování nevhodných znaků (lomítka, opakované mezery, mezery na konci či

počátku). U proměnných založených na číselníku bylo třeba v některých případech nahradit chybějící hodnotu kódem pro chybějící hodnotu z číselníku. Číselníky obsahují možnost neznámé, či chybějící hodnoty, např. *Neuvedeno*, Bez *informací* apod.

Z existujících proměnných byly odvozeny nové proměnné vhodné pro další analýzu, šlo zejména o extrakci dne v týdnu, měsíce a roku z data úrazu a doplnění informace o státních svátcích.

### 2.2.3. Připojení číselníků

Zdrojová data obsahovala číselníkové hodnoty, jež byly příslibem snadného zpracování, ale podle informací od poskytovatele dat nebylo možné číselníkové kódy použít. V určité fázi procesu tvorby a prvotní úpravy dat u třetích stran došlo ke ztrátě úvodní nuly kódu ve formátu 0XXX. Proto bylo nutné pracovat hlavně s textovými hodnotami a kódy použít pouze pro případné doplnění. Originální hodnoty nejsou pro reportování či analýzu vhodné, protože pocházejí z různých úrovní číselníku a hodnoty uváděné na nejnižších úrovních jsou příliš roztříštěné. V analýze je vhodné pracovat s nejvyššími úrovněmi číselníků. Textové hodnoty byly také v některých případech upravovány a neodpovídaly číselníkovým textům.

Z uvedených důvodů bylo nutné z číselníku připojit kód odpovídající danému textu a nadřazené úrovně. Před připojením byly chybné texty uživatelsky upraveny na číselníkové texty, např. text v datech *nebezpečné látky* byl nahrazen odpovídající číselníkovým textem *Nebezpečné látky a přípravky, radioaktivní látky, biologické látky – nespecifikováno*. Při zpracování dat z dalších let postižených stejnými problémy bude nutné opravy textů doplnit. K opraveným textům byl doplněn příslušný kód a nadřazené úrovně, připojovaly se tedy kódy podle textů. V případech, kdy textová hodnota nebyla k dispozici, nebo se ji nepodařilo opravit, bylo vyzkoušeno opačné připojení textu podle kódu. Výsledné doplněné hodnoty byly zkombinovány s upřednostněním hodnot doplněných podle textu. Tímto postupem se dosáhlo vysoké vyplněnosti číselníkových hodnot.

### 2.2.4. Závěrečné úpravy

Výsledné proměnné byly doplněny o metadata (formáty, typy proměnných, popisky), proměnné byly také uspořádány do logického pořadí. Pro snadné zpracování v portálu obsahují výsledná data číselníkový kód i textovou hodnotu.

## 2.3. Seznam proměnných ve výstupním souboru

Výsledný soubor se nazývá ***PU_SUIP.sav*** a obsahuje následující proměnné. V seznamu jsou uvedeny číselné hodnoty v lomených závorkách <> a kategorie v hranatých závorkách []. Kde to počet kategorií dovolil, jsou kategorie vypsány, u uspořádaných kategorií je vypsána první a poslední hodnota oddělená znakem pomlčky.

– Identifikace úrazu
  - datum_urazu – datum úrazu <1.1.2014;31.12.219>
  - zdroj_dat – instituce poskytující data [SUIP; ČBÚ]
  - rok_urazu – rok úrazu <2014;2019>
  - mesic_urazu – měsíc úrazu <1;12>
  - den_tydne_urazu - <1;7>
  - svatek- příznak svátku [0;1]
  - den_pracovni- příznak pracovního dne [0;1]

– Informace o osobě s úrazem
  - pohlavi – pohlaví osoby [muž; žena]
  - vek – věk osoby <15;91>
  - vek_kat – věkové kategorie [< 18 ;18 – 19; 20 – 24; 25 – 29; 30 – 34; 35 – 39; 40 – 44; 45 – 49; 50 – 54; 55 – 59; 60+]
  - CZNACE_KOD – kód odvětví ekonomické činnosti zaměstnavatele dle číselníku
  - CZNACE – textová hodnota odvětví ekonomické činnosti zaměstnavatele dle číselníku
  - CZNACE_nadrazene_KOD – kód nejvyšší úrovně ekonomické činnosti zaměstnavatele dle číselníku [0;A-U]
  - CZNACE_nadrazene – textová hodnota nejvyšší úrovně ekonomické činnosti zaměstnavatele dle číselníku
  - KRAJ_KOD – NUTS 3 kód kraje zaměstnavatele [CZ010 – CZ080; CZ099]
  - KRAJ – název kraje zaměstnavatele [Praha – Moravskoslezský]
  - OKRES_KOD – LAU 1 kód okresu zaměstnavatele [CZ0100 – CZ0806; CZ9999]
  - OKRES – název okresu zaměstnavatele
  - KZAM – kód profese zaměstnance dle úrovně číselníku (používáno do roku 2018)
  - KZAM_nadrazene_KOD – kód profese zaměstnance dle nejvyšší úrovně číselníku (používáno do roku 2018) [1000;10000–90000]
  - KZAM_nadrazene_– textová hodnota profese zaměstnance dle nejvyšší úrovně číselníku (používáno do roku 2018)
  - ISCO – kód profese zaměstnance dle číselníku (používáno od roku 2019)
  - ISCO_nadrazene_KOD – kód profese zaměstnance dle nejvyšší úrovně číselníku (používáno od roku 2019) [0;10000–90000]
  - ISCO_nadrazene_nazev – textová hodnota profese zaměstnance dle nejvyšší úrovně číselníku (používáno od roku 2019)
– Specifikace úrazu
  - druh_urazu_KOD – kód druhu úrazu dle číselníku
  - druh_urazu – textová hodnota druhu úrazu dle číselníku
  - cinnost_pri_urazu_KOD – kód druhu úrazu dle číselníku
  - cinnost_pri_urazu – textová hodnota činnosti při úrazu dle číselníku
  - cinnost_pri_urazu_nadrazena1_KOD – kód činnosti dle 2. nejvyšší úrovně číselníku
  - cinnost_pri_urazu_nadrazena1 – textová hodnota činnosti dle 2. nejvyšší úrovně číselníku
  - cinnost_pri_urazu_nadrazena2_KOD – kód činnosti dle nejvyšší úrovně číselníku
  - cinnost_pri_urazu_nadrazena2 – textová hodnota činnosti dle nejvyšší úrovně číselníku [0;1000–4000]
  - zdroj_urazu_KOD – kód věci nebo jevu, který způsobil úraz, dle číselníku
  - zdroj_urazu – textová hodnota věci nebo jevu, který způsobil úraz, dle číselníku
  - zdroj_urazu_nadrazeny_KOD – kód věci nebo jevu, který způsobil úraz, dle nejvyšší úrovně číselníku
  - zdroj_urazu_nadrazeny – textová hodnota věci nebo jevu, který způsobil úraz, dle nejvyšší úrovně číselníku

- misto_urazu_KOD – kód místa úrazu dle číselníku
- misto_urazu – textová hodnota místa úrazu dle číselníku
- misto_urazu_nadrazene_KOD – kód místa úrazu dle nejvyšší úrovně číselníku
- misto_urazu_nadrazene – textová hodnota místa úrazu dle nejvyšší úrovně číselníku
- druh_zraneni_KOD – kód druhu zranění dle číselníku
- druh_zraneni – textová hodnota druhu zranění dle číselníku
- druh_zraneni_nadrazeny_KOD – kód druhu zranění dle nejvyšší úrovně číselníku
- druh_zraneni_nadrazeny – textová hodnota druhu zranění dle nejvyšší úrovně číselníku
- cast_tela_KOD – kód části těla dle číselníku
- cast_tela – textová hodnota části těla dle číselníku
- cast_tela_nadrazena1_KOD – kód části těla dle 2. nejvyšší úrovně číselníku
- cast_tela_nadrazena1 – textová hodnota části těla dle 2. nejvyšší úrovně číselníku
- cast_tela_nadrazena2_KOD – kód části těla dle nejvyšší úrovně číselníku [0;1000–7000;9900]
- cast_tela_nadrazena2 – textová hodnota části těla dle nejvyšší úrovně číselníku

# 3. Invalidita

Data o invaliditě pocházejí z let 2014–2018, jejich poskytovatelem je **Česká správa sociálního zabezpečení**. Vstupní data jsou ve formátu ***MS Excel*** *.xlsx*. Jeden soubor obsahuje údaje z jednoho roku a pro jeden stupeň invalidity. Stupně jsou 3. Data jsou v každém souboru umístěna na jednotlivých listech. List je určen kombinací pohlaví a skupiny diagnóz podle *Mezinárodní klasifikace nemocí* (*MKN-10*). Každý list obsahuje počet nových invalidních důchodů pro jednotlivé diagnózy a věkové skupiny. Soubory obsahují i součtové listy, které nebyly požity, jelikož se dají ze základních dat dopočítat sečtením.

Jako doplňková informace byl použit soubor s počtem obyvatel *Obyvatele_CR.sav* (viz odstavec 6.1).

## 3.1. Datové manipulace

Výsledný datový zdroj obsahuje spojená data za jednotlivé roky a stupně invalidity. Prvním krokem je proto sloučení dat do jednoho souboru. Pro další analýzy je nutné mít data v tzv. dlouhém formátu, kdy sada proměnných definuje charakter datových hodnot, které jsou uspořádány pod sebou. Zde jsou definiční proměnné *Rok*, *Stupeň*, *Skupina* + D*iagnóza*, *Pohlaví* a *Věková kategorie*, jejich kombinace tvoří jeden řádek, datová hodnota je *počet invalidních důchodů*. Po sjednocení dat proto následovala změna jejich struktury. V dalším kroku byly sloučeny řídce obsazené diagnózy a připojen příslušný počet obyvatel. Výsledkem je soubor ve formátu *.sav*.

### 3.1.1. Spojení

Vstupní data tvoří jednotlivé soubory a listy v nich. 5 let a 3 stupně invalidity definují 15 souborů. Jeden list obsahuje kombinaci *skupiny diagnóz* a *pohlaví*. V souborech se diagnózy člení do 20 skupin, z nichž poslední je skupina *Neuvedeno*. Na každém listě jsou hodnoty uspořádány v tabulce s řádky definovanými *diagnózami* a sloupci *věkovými kategoriemi*. Z každého sešitu je nutné načíst 40 tabulek, což tvoří celkem 600 tabulek. Poskytovatel dat zaručuje stejnou strukturu dat v souborech a na listech.

Při spojování tabulek je nutné načíst kromě samotných dat i název zdrojového souboru, protože ten určuje *rok* a *stupeň invalidity*, a název listu, protože ten určuje skupinu *diagnóz* a *pohlaví*. Spojení bylo realizováno makrem v jazyce ***Visual Basic***. Makro zpracovalo všechny soubory a v nich všechny relevantní listy, získalo jejich názvy a data a vše zkopírovalo do výsledného soboru *.xlsx* na jeden list. Makro vytváří i kontrolní list, kde lze ověřit, zda struktura tabulek je na všech zdrojových listech stejná. To bylo pro zpracovávaná data splněno.

### 3.1.2. Restrukturalizace a rekategorizace

Ve spojeném souboru jsou počty invalidních důchodů pro jednotlivé věkové kategorie v jednotlivých sloupcích. Ve výsledné struktuře musejí být hodnoty počtů důchodů v jednom sloupci s identifikací věkové kategorie ve speciální proměnné. Restrukturalizace a všechny následné datové operace byly provedeny v programu ***SSPS Modeler***. Následovala extrakce *Roku*, *stupně invalidity*, *skupiny diagnóz* a *pohlaví* z názvu souborů a listů.

Velký počet diagnóz má za následek nízkou nebo dokonce nulovou obsazenost většiny z nich. Aby byla analýza smysluplná, byly diagnózy s celkovým počtem případů bez ohledu na stupeň a rok menším než 10 sloučeny ve své skupiny do jedné diagnózy. Výsledkem sloučení je nová proměnná, která pro málo obsazené diagnózy nabývá hodnoty *XY_99*, kde *XY* je číslo skupiny diagnóz. Pro ostatní diagnózy kopíruje původní hodnotu. Při slučování nebyla data fyzicky agregována, ale pouze byla vytvořena proměnná pro analýzu a reportování. Operace tedy nevedla ke ztrátě dat.

### 3.1.3. Připojení počtu obyvatel a závěrečné úpravy

Pro konstrukci relativních údajů je nutné, aby data obsahovala počty obyvatel v jednotlivých letech a věkových kategoriích. Zdrojový soubor *obyvatele_CR.sav* obsahuje počty pro kombinaci roku a pohlaví. Počty obyvatel byly sečteny do věkových kategorií podle invalidity a připojeny k datům o invaliditě. Z dat byly odstraněny diagnózy s nulovým počtem případů a data vhodně uspořádána. Výsledkem je datový soubor ve formátu *.sav*.

## 3.2. Seznam proměnných ve výstupním souboru

Výsledný soubor se nazývá ***INV_CSSZ.sav*** a obsahuje následující proměnné. V seznamu jsou uvedeny číselné hodnoty v lomených závorkách <> a kategorie v hranatých závorkách []. Kde to počet kategorií dovolil, jsou kategorie vypsány, u uspořádaných kategorií je vypsána první a poslední hodnota oddělená znakem pomlčky.

- rok – rok přiznání invalidního důchodu <2014;2018>
- stupen_invalidity – stupeň invalidity [1;2;3]
- skupina_skupina – skupina diagnóz dle číselníku [1 – 20]

- skupina_slovne – textový název skupiny diagnóz dle číselníku
- skupina_kody – rozsah kódů diagnóz nacházejících se v dané skupině diagnóz, hodnoty dle číselníku
- kod_diagnozy_KAT – hodnota kódu diagnózy dle číselníku se sloučením málo četných diagnóz do jedné, sloučené kategorie: [01_99 – 20_99]
- kod_diagnozy– hodnota kódu diagnózy dle číselníku, nesloučené hodnoty
- pohlavi – pohlaví důchodce [muž, žena]
- vek_KAT– věkové kategorie důchodce [0 – 19; 20 – 24; 25 – 29; 30 – 34; 35 – 39; 40 – 44; 45 – 49; 50 – 54; 55 – 59; 60 – 64; 60+]
- pocet – počet přiznaných důchodů
- pocet_obyvatel – střední stav obyvatelstva (k 1.7.)

# 4. Dočasná pracovní neschopnost

Data o dočasné pracovní neschopnosti (DPN) pocházejí z let 2014 – 2018, jejich poskytovatelem je **Česká správa sociálního zabezpečení**. Vstupní data jsou ve formátu *.xlsx*. Jeden soubor obsahuje údaje za jeden měsíc Data jsou v každém souboru umístěna na jednom listě v několika tabulkách, které se liší důvodem ukončení pracovní neschopnosti. Uvnitř tabulek je rozlišeno pohlaví a interval délky pracovní neschopnosti. Hodnoty tvoří počet ukončených pracovních neschopností a celkový počet prostonaných dní. Listy obsahují i součtové tabulky, které nebyly požity, jelikož se dají ze základních dat dopočítat sečtením.

## 4.1. Datové manipulace

Výsledný datový zdroj obsahuje spojená data za celé období. Prvním krokem je proto sloučení dat do jednoho souboru. Pro další analýzy je nutné mít data v tzv. dlouhém formátu, kdy sada proměnných definuje charakter datových hodnot, které jsou uspořádány pod sebou. Zde jsou definiční proměnné *rok*, *měsíc*, *důvod ukončení pracovní neschopnosti*, *interval jejího trvání* a *pohlaví*, jejich kombinace tvoří jeden řádek, datové hodnoty jsou *počet ukončených pracovních neschopností* a *počet prostonaných dní*. Dalším krokem úprav byla proto změna struktury dat. Posledním krokem bylo uspořádání dat a doplnění metadat (popisky, formáty). Výsledkem je soubor ve formátu sav.

### 4.1.1. Spojení

Vstupní data tvoří 62 jednotlivých souborů. Z nich je nutné načíst data z 5 tabulek lišících se způsobem ukončení pracovní neschopnosti. Celkem se načítá 190 tabulek. Poskytovatel dat zaručuje stejnou strukturu dat v souborech a na listech.

Při spojování tabulek je nutné načíst kromě samotných dat i název zdrojového souboru, který určuje rok a měsíc, a nadpis tabulky určující způsob ukončení. Načtení a sloučení bylo provedeno v programu ***SPSS Modeler***.

### 4.1.2. Restrukturalizace a závěrečné úpravy

Ve spojeném jsou souboru počty pro odlišné pohlaví v jednotlivých sloupcích. Ve výsledné struktuře musejí být hodnoty počtů v jednom sloupci s identifikací pohlaví ve speciální

proměnné. Restrukturalizace, doplnění metadat a uložení do výsledného souboru ve formátu sav bylo provedeno opět v programu ***SPSS Modeler***.

### 4.2. Seznam proměnných ve výstupním souboru

Výsledný soubor se nazývá ***DPN_CSSZ.sav*** a obsahuje následující proměnné. V seznamu jsou uvedeny číselné hodnoty v lomených závorkách <> a kategorie v hranatých závorkách []. Kde to počet kategorií dovolil, jsou kategorie vypsány, u uspořádaných kategorií je vypsána první a poslední hodnota oddělená znakem pomlčky.

- rok – rok ukončení pracovní neschopnosti <2014;2018>
- mesic – měsíc ukončení pracovní neschopnosti [1 – 12]
- duvod_pracovni_neschopnosti – důvod, ze kterého pracovní neschopnost vznikla [Nemoc; Úraz; Pracovní úraz; Úraz zaviněný jinou osobou; Alkohol / Omamné látky]
- interval – kód intervalu délky pracovní neschopnosti [1 – 10]
- interval_TXT – textová hodnota intervalu délky pracovní neschopnosti [1 – 3 dní; 4 – 14 dní; 15 – 21 dní; 22 – 30 dní; 31 – 60 dní; 61 – 90 dní; 91 – 180 dní; 181 – 270 dní; 271 – 365 dní; > 365 dní]
- skupina_kody – rozsah kódů diagnóz ve skupině diagnóz, hodnoty dle číselníku NKM-10
- pohlavi – pohlaví důchodce [muž, žena]
- vek_KAT– věkové kategorie důchodce kategorie [0 – 19; 20 – 24; 25 – 29; 30 – 34; 35 – 39; 40 – 44; 45 – 49; 50 – 54; 55 – 59; 60 – 64; 60+]
- počet_PN – počet ukončených pracovních neschopností
- prostonane_dny – celkový počet dní trvání pracovních neschopností

## 5. Pracovní neschopnost

Data o invaliditě pocházejí z let 2014–2018, zdrojové soubory jsou ve formátech *.csv* a *.sav*. Data poskytnul **Ústav zdravotnických informací a statistiky** a pocházejí z *Registru pracovní neschopnosti*. Řádek v datech představuje jednu pracovní neschopnost (PN).

### 5.1. Datové manipulace

Formát dat byl prakticky vyhovující pro analýzu. Datové zdroje bylo třeba pouze spojit do jednoho souboru. poté následovalo pouze přejmenování technických proměnných a uložení souboru do formátu *.sav*.

### 5.2. Seznam proměnných ve výstupním souboru

Výsledný soubor se nazývá ***PN_UZIS.sav*** a obsahuje následující proměnné. V seznamu jsou uvedeny číselné hodnoty v lomených závorkách <> a kategorie v hranatých závorkách []. Kde to počet kategorií dovolil, jsou kategorie vypsány, u uspořádaných kategorií je vypsána první a poslední hodnota oddělená znakem pomlčky.

- pohlavi – pohlaví [muž;žena]
- vek – věk v letech <14;95>

- vekova_kategorie – věkové kategorie [< 20; 20 – 24; 25 – 29; 30 – 34; 35 – 39; 40 – 44; 45 – 49; 50 – 54; 55 – 59; 60 – 64; 60+]
- rok_narozeni – poslední dvojčíslí roku narození <22;99>
- mesic_narozeni – měsíc narození v roce [leden – prosinec]
- rok_pocatku_PN – rok, kdy PN vznikla, <2012;2017>, roky 2012 a 2013 zastoupeny zanedbatelně a vyloučeny z dalších analýz
- mesic_pocatku_PN – měsíc počátku PN v roce [leden –prosinec]
- den_pocatku_PN – den počátku PN <1;31>
- rok_ukonceni_PN – rok, kdy byla PN ukončena, <2014;2017
- mesic_ukonceni_PN – měsíc ukončení PN v roce [leden – prosinec]
- den_ukonceni_PN – den ukončení PN <1;31>
- delka_PN – délka PN ve dnech <1;996>
- interval_PN – kategorie délky PN
- zpusob_ukonceni_PN_kod – kód události, která ukončila PN [0 – 9]
- zpusob_ukonceni_PN_nazev – název události, která ukončila PN [důchod I, Ič; důchod starobní; jiný způsob; mateřská dovolená; nástup do lázní; práce schopen; předání mimo okres; storno potvrzení PN; ukončení lékařem správy; vyčerpání dávek]
- diagnoza_kod – kód diagnózy vedoucí k PN, dle číselníku NKM-10
- diagnoza_skup_kod_int – rozsah kódů diagnóz ve skupině diagnóz, hodnoty dle číselníku NKM-10
- diagnoza_skup – textový název skupiny diagnóz dle číselníku
- diagnoza_skup_NUM – kód skupiny diagnóz dle číselníku [1 – 19;99]
- druh_urazu – druh PN [nemoc; karanténa; karanténa – epidemie]
- KZAM_kod – kód profese zaměstnance dle úrovně číselníku
- KZAM_nazev – název profese zaměstnance dle úrovně číselníku
- typ_zuctovatele – zúčtovatel PN [OSVČ; Zaměstnanec]
- okres_PN_kod – LAU 1 kód okresu místa pobytu v době PN
- okres_PN – název okresu místa pobytu v době místa pobytu v době PN
- kraj_PN – kód kraje místa pobytu v době PN [1 – 14]
- kraj_text_PN – název kraje místa pobytu v době PN [Praha – Moravskoslezský]
- okres_zamestnavatele_kod – LAU 1 kód okresu zaměstnavatele
- okres_zamestnavatele – název okresu zaměstnavatele
- kraj_zamestnavatele – kód kraje zaměstnavatele PN [1 – 14]
- kraj_text_zamestnavatele – název kraje zaměstnavatele [Praha – Moravskoslezský]
- okres_ordinace_kod – LAU 1 kód okresu ordinace
- okres_ordinace – název okresu ordinace
- kraj_ordinace – kód kraje ordinace [1 – 14]
- kraj_text_ordinace – název kraje ordinace [Praha – Moravskoslezský]

# 6. Doplňkové soubory

Pro výpočet relativních ukazatelů je nutné pracovat s počty na národní úrovni. Uvedená dat má smysl vztahovat k celkovému počtu obyvatel, respektive k počtu zaměstnanců.

Globální data by měla být ve stejné struktuře, jako jsou analyzovaná data. V praxi je možná struktura určena poskytovatelem dat, což je **Český statistický úřad**.

## 6.1. Počty obyvatel

Zdrojem dat o počtu obyvatel je *Věkové složení obyvatelstva k 1.7.* za jednotlivé roky publikované ČSÚ pod kódy produktu *130064-15* až *130064-19*. Soubory byly spojeny v programu ***SPSS Modeler***, vybaveny metadaty a uloženy do souboru ***Obyvatele_CR.sav***.

Proměnné:

– ROK– Rok <2014;2018>
– POHL – Pohlaví [Muž;Žena]
– VEK – Věk v letech <0;100>
– Počet – počet obyvatel k 1.7.

## 6.2. Počty zaměstnanců

Zdrojem dat o počtu obyvatel jsou data z *Veřejné databáze ČSÚ*. Uživatelské rozhraní je přístupné pod odkazem https://vdb.czso.cz/vdbvo2/faces/cs/index.jsf?page=uziv-dotaz#. Pracuje se s ukazatelem *//zaměstnání//odvětví ekonomické činnosti [21]* (***Zaměstnanost, nezaměstnanost/Zaměstnání/odvětví ekonomické činnosti [21]***) v členění za kraje a v letech 1993–2019. Vygenerovaný soubor *.xlsx*. byl v programu ***SPSS Modeler*** restrukturalizován do dlouhého formátu, doplněn metadaty a uložen do formátu *.sav*. Vznikl soubor ***Pocet_zam_ROK_NACE_KRAJ_VerDtb.sav***.

Proměnné:

ROK – Rok <1993;2019>
KRAJ_KOD – NUTS 3 kód kraje zaměstnavatele [CZ010 – CZ080; CZ099]
CZNACE_nadrazene_KOD – kód nejvyšší úrovně ekonomické činnosti [A – S]
Pocet_zam – počet zaměstnanců v tisících

## 6.3. Číselníky

Některé proměnné jsou v uvedených datových souborech vytvářeny na základě číselníků. Číselníky je v datovém procesu vhodné použít, ke kontrole přípustných hodnot, kódů i textů, a u některých číselníků k připojení nadřazených kategorií. Poskytovatelem číselníků je **Český statistický úřad** a **Státní úřad inspekce práce**.

– **Český statistický úřad**
  - Klasifikace ekonomických činností (CZ-NACE)
  - Klasifikace zaměstnání (CZ-ISCO)
  - Klasifikace územních statistických jednotek (CZ-NUTS)
  - Mezinárodní statistická klasifikace nemocí a přidružených zdravotních problémů (MKN-10)
– **Státní úřad inspekce práce**
  - Druh úrazu
  - Činnosti při úrazu
  - Zdroj úrazu
  - Místo úrazu
  - Druh zranění

- Zraněná část těla

# 7. Doporučení na změnu datových zdrojů

Protože byla úprava dat z výchozích formátů do stavu vhodného k analýze či reportu náročná a u některých datových zdrojů se informace podařilo využít jen částečně (pracovní úrazy), bylo by vhodné strukturu výchozích dat upravit a změnit možnosti přístupu k nim.

Všechna data by měla být ideálně uložena v relačních databázích a měla by splňovat následující požadavky:

a) alespoň 1. normální forma,
b) založení na číselnících a hvězdicové schéma,
c) dlouhý formát,
d) jednotnost.
e) neodvoditelnost,
f) dostupnost

a) 1. normální forma je splněna, pokud jeden datový element, např. buňka v databázové tabulce nebo v tabulkovém souboru, případně hodnota oddělená oddělovači v textovém souboru, obsahuje jen jednu hodnotu. Příkladem porušení je spojení kódu a textové hodnoty do jednoho údaje. Takové porušení lze někdy vyřešit, ale za cenu rozsáhlých datových manipulací. Závažnějším porušením je, když jedno pole v databázi obsahuje významově odlišné hodnoty, aniž by to bylo upřesněno jinými poli. Například, pokud pole *OKRES*, obsahuje v pro některé případy okres sídla zaměstnavatele a pro jiné místo trvalého pobytu zaměstnance. Takové porušení lze vyřešit jen zřídka, vyžaduje to dodatečnou informaci o významu jednotlivých hodnot.

b) Proměnné, které mohou nabýt jen specifikovaných hodnot, by měly být založeny na číselnících. Tabulka s fakty obsahuje pouze číselníkovou hodnotu (kód). Textový popis kódu je obsažen v oddělené tabulce a mezi tabulkami je jasně definované propojení. Výhodou je krom menší velikosti databáze hlavně přítomnost jen povolených hodnot. Případná úprava nebo oprava textové hodnoty probíhá jen na jednom místě. Pro záměr analyzovat vývoj v delším období je nutné používat číselník pro danou proměnnou celé sledované období. Při přechodu na nový číselník je nutné definovat jednoznačné přiřazení hodnot starého a nového číselníku.

c) Dlouhý formát (viz 3.1nebo 4.1.) znamená, že jedna proměnná je obsažena v jednom poli. Čeho se hodnota proměnné týká, je uvedeno v zvláštních polích. Příkladem porušení jsou sloupce s proměnná *DPN* je ve dvou sloupcích pro muže a ženy. Převod dat do dlouhého formátu je řešitelný, ale vyžaduje další datové manipulace. V uvedeném příkladě bude dlouhý formát obsahovat jedno pole s proměnnou *DPN* s hodnotami původních polí pod sebou a nové pole rozlišující, kterého pohlaví se hodnota v *DPN* týká.

d) Data by měla být uložena v jedné tabulce, resp. souboru. Data z různých oblastí nebo let by neměla být v různých tabulkách či souborech. V komplexní tabulce je určeno, které oblasti se konkrétní údaj týká v příslušném poli. Sjednocení dat z více souborů do jednoho s doplněním významu hodnot je možné, ale vyžaduje další datové manipulace, jejichž množství může dosáhnout tak vysokých hodnot, že datový proces se stává nepřijatelně těžkopádným.

e) Uložené hodnoty by měly být neodvoditelné z jiných hodnot. Pokud se dá hodnota odvodit z jiných hodnot, není zaručena konzistence dat. Typickým příkladem je uvádění součtu z jiných hodnot jako samostatná hodnota. Hodnota součtu pak nemusí odpovídat skutečnému součtu jednotlivých hodnot. Původní hodnoty se mohly změnit nebo součet byl spočten chybně. Odvozená hodnota také zbytečně zvětšuje velikost dat. Odvozené hodnoty se správně počítají až v průběhu reportování nebo analýzy, tím je zaručeno, že jsou založeny na aktuálních hodnotách. Snadným řešením je ignorování odvozených hodnot, ale je třeba informace, které hodnoty jsou primární a které odvozené.

f) V ideální variantě by měla být data přístupná přímo pro tvůrce reportů nebo analýz s příslušným oprávněním. Absence exportu dat do jiného formátu (*.csv*, *.xlsx*., *.sav*) by eliminovala riziko neúplnosti dat způsobené nevhodným kódováním či automatickou změnou formátu exportovaných dat v některých aplikacích. Analytik může také využívat všech dat a rychle reagovat na změněné požadavky na analýzu. Přímý přístup lze snáze zabezpečit než fyzický přenos datových souborů nebo přenos přes internet.

Uvedené požadavky lze krom bodu f) splnit i v jiné než databázové formě, nabízí se formát *.xlsx* nebo *.sav*.