Procvičování 9a

1. Pohrajeme si ještě s těmi jmény, co jsme s nimi pracovali posledně. Tentokrát vám nebudu diktovat, jaké objekty máte v jakém bodě vytvářet a jak je pojmenovat. Spíš po vás budu chtít něco zjistit, případně zobrazit, a jaké objekty si k pomoci vytvoříte nechám na vás. Taky to bude trochu o práci s nápovědou. Pokud dataset s frekvencemi jmen českých občanů *jmena.txt* nemáte importovaný z minula, importujte jej do **R**. Najdete jej ve studijních materiálech, nebo ji můžete načíst přímo z `http://www.sci.muni.cz/~syrovat/jmena.txt`.Tabulka obsahuje v řádcích všechna jednoslovná křestní jména českých obyvatel, ve sloupcích pak jejich frekvenci v jenotlivých ročnících od r. 1896 do 2013.

2. Celý dataset obsahuje hromadu balastních jmen. Mám na mysli jména, jež jsou velmi vzácná, aby se s nimi dalo něco dělat. Jedná se o jména cizinců, či o špatně zadaná jména. Odstraníme tedy vzácná jména. Nejprve si ale namalujeme dotchart, abychom si udělali obrázek o frekvencích jmen. Dotchart (Cleveland dot plot) je jednou z grafických možností pro posouzení rozložení hodnot. Spočívá ve vynešení vlastních hodnot na osu X podle jejich pořadí (Y). Pro přehlednost můžeme hodnoty seřadit. Můžeme použít funkci `dotchart()`, která ale není moc neflexibilní. Raději si dotchart vytvoříme sami: zobrazte seřazené frekvence jmen na ose X a jejich pořadí na ose Y.

3. Je vidět, že opravdu velká většina z oněch 16 tisíc jmen je velmi vzácná a teprve u posledního 1-2 tisíce jmen jejich frekvence narůstá, můžeme si tedy dovolit jich většinu odstranit. Kolik ale? Hranice by měla ležet někde v ohybu té křivky, kterou tvoří body v našem dotchartu. Já jsem zvolil arbitrární kritérium 0,5%. "Odstraňujme vzácná jména postupně od těch nejvzácnějších a zastavme s odstraňováním těsně předtím, než bychom dosáhli 0,5% odstraněných jedinců." Tak docílíme toho, že nám v datasetu zůstane naprostá většina dat (alespoň 99,5%), a zároveň ho vyčistíme od balastu. Vypočítal jsem, že to odpovídá všem jménům s frekvencí menší nebo rovnou 33. To můžete ověřit. Zjistěte, kolik jmen a kolik jedinců nám z datasetu tímto postupem vypadne, kolik tito jedinci tvoří procent přesně, a z datasetu je odstraňte. Od teď budeme pracovat už jen s tímto menším a čistším datasetem.

4. Už víme, že nejběžnějším jménem je Jiří. Jak je to s dalšími jmény, která začínají na "J"? Zjistěte 10 nejběžnějších jmen začínajících písmenem "J". K tomu budete potřebovat nějakou funkci, která dokáže v textového řetězce vyextrahovat první písmenko. Tu najděte. Jen napovím, že pro řetězec se používá termín "string".

5. Teď, když si umíme vytáhnout jména stejného začátečního písmene, bychom se mohli podívat, jak si stojí naše jména s frekvencí v množině jmen stejného začátečního písmene: Zjistěte pořadí Vašeho jména mezi jmény začínajícími stejným písmenem. (Pokud Vaše jméno vypadlo v bodu 2, zjistěte pořadí nějakého jiného oblíbeného jména).

6. Protože data jsou strukturována podle roku narození, můžeme zjistit něco o trendech v používání jmen během posledních desetiletí. Namalujte sloupcový graf relativního zastoupení Vašeho jména v populaci v jednotlivých letech. Výsledkem má být graf se sloupečky, kde počet sloupečků odpovídá počtu let, pro které jsou frekvence jmen v datasetu zaznamenané, a výška sloupců relativnímu zastoupení daného jména mezi jedinci narozenými v patřičném roce. Namalujte stejný graf pro tři různá jména, u kterých byste čekali odlišný vývoj. Sloupečky ve sloupečkovém grafu se označují termínem "bar".