Dobrovolný úkol - Přesnost předvoleních průzkům v USA

# Úvod

V rámci našeho dobrovolného úkolu se podíváme na téma, které se pravidelně stává populární každých pár let: přesnost předvolebních průzkumů. Konkrétně budeme pracovat s daty poskytnutými organizací [FiveThirtyEight](https://fivethirtyeight.com/). FiveThirtyEight je americkou organizací zabývající se datovou žurnalistikou, a to zejména tématy politiky a sportu. Spravuje také jeden z nejpopulárnějších agregátorů průzkumů veřejného mínění s politickou tématikou.

My budeme pracovat s daty o jednotlivých předvolebních průzkumech v USA za posledních zhruba 20 let.

# Zadání

V rámci domácího úkolů se pokuste zodpovědět následující otázku:

*Liší se přesnost předvolebních průzkumů (proměnná* ***error****) v závislosti na metodologii sběru dat (****methodology****)?*

Data k analýze si stáhněte pomocí následujícího kódu:

polls = read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/pollster-ratings/raw-polls.csv")  
polls = subset(polls, methodology == c("IVR", "Live Phone", "Online"))

(Z dat si také pro zjednodušení vyfiltrujeme pouze určité kategorie methodology. Za malé bezvýznamné plus se můžete podívat na celková data a zamyslet se nad tím, proč jsem vybral jen tyhle tři.)

Popis dat najdete v tabulce zde: <https://github.com/fivethirtyeight/data/tree/master/pollster-ratings>

Pro inspiraci a pro kontext si můžete přečíst následující článek: [The Death Of Polling Is Greatly Exaggerated](https://fivethirtyeight.com/features/the-death-of-polling-is-greatly-exaggerated/).

# Co by měla vaše práce obsahovat

## Diskuze o výběru relevantních proměnných

Předtím, než se pustíte do modelování, zamyslete se nad tím, jestli by náš model měl obsahovat ještě nějaké další proměnné. Jsou nějaké jiné proměnné, pro které bychom měli kontrolovat, abychom získali lepší/přesnější odpověď na naši otázku? Jsou nějaké proměnné, kterým bychom se naopak měli vyhnout? A může se se vztah mezi závislou a nezávislou proměnnou lišit v závislosti na třetí proměnné (tj. vyplatí se přidat do našeho modelu nějakou interakci)?

Věnujte výběru proměnných jeden až dva odstavce.

## Základní popis proměnných, se kterými pracujeme

Až si vyberete své proměnné, krátce si je prohlédněte, abyste získali představu o datech, se kterými pracujete. Podívejte se na jejich rozdělení a popisné statistiky, a poznamenejte si vše, co vám přijde zajímavé nebo potenciálně důležité. Např. jaká je průměrná chyba předvolebních průzkumů a jak vypadá jejich rozdělení? Kolik průzkumů využívá každou z metod sběru dat? Podívejte se i na další relevantní proměnné, které jste vybrali v předchozím kroku.

Věnujte téhle části znovu jeden až dva odstavce. Nebojte se připojit graf, pokud vám přijde relevantní

## Vytvoření a interpretace modelu

Na základě předchozích kroků vytvořte model, pomocí kterého budeme moct zodpovědět naši výzkumnou otázku.

Zkuste model interpretovat pomocí tabulky regresních koeficientů, i pomocí grafů marginálních efektů (balíček ggeffects, který jste si ukazovali s Járou). Nezapomeňte, že vizuální interpretace je často jednoduší! Interpretujte jak bodové odhady (regresní koeficienty nebo predikované hodnoty), tak i míru nejistoty s nimi spojenou (standardní chyba nebo intervaly spolehlivosti).

Nebojte se v této části více rozepsat a napsat vše, co vás k interpretaci napadne. Přidejte grafy marginálních efektů, pokud vám přijdou relevantní.

## Hodnocení modelu a diagnostika

Zhodnoťte model jako celek. Začněte tím, jak moc náš model sedí na data. V jakém ohledu je tato informace relevantní vzhledem k naší výzkumné otázce? Dále se podívejte na předpoklady modelu.

Zamyslete se nad všemi předpoklady, které s vaším modelem souvisí. Tam, kde to bude možné, je zkontrolujte pomocí diagnostických grafů. Klidně si s daty pohrajte (např. pokud vám bude připadat, že nějaké pozorování je přehnaně vlivné, zkuste model spočítat bez něj). Nestresujte se ovšem, pokud váš model nebude vypadat učebnicově a splňovat předpoklady do puntíku. Zaměřte se se spíše na jejich zhodnocení a kriticky se zamyslete, k čemu může vést jejich porušení.

## Celkové zhodnocení

Na základě všeho předchozího odpovězte naši výzkumnou otázku. Zamyslete se nad tím, jak moc jste se svojí odpovědí spokojení. Získali jste jednoznačnou odpověď nebo zůstávají ještě nějaké dílčí otázky. Jak moc důvěřujete svému modelu? Napadá vás něco dalšího, čím by se dal model potencionálně vylepšit (např. nechybí nám v datech nějaká důležité proměnná? Má model nějakou slabinu, kterou byste chtěli podchytit, ale nenapadá vás jak?)?

# Formát a termín odevzdání

Text, obsahující všechny výše zmíněné body, odevzdejte jako Word nebo PDF soubor. Přiložte i R skript, který bude obsahovat postup analýzy. Celý skript by měl být samostatně funkční, tzn. mělo by možné ho spustit celý najedou a replikovat všechny vaše výsledky. Úkoly posílejte na [alesvomacka@seznam.cz](mailto:alesvomacka@seznam.cz)