Prezentace balíčku Naniar

Hana Oulehlová, Martin Štýber

K čemu slouží balíček Naniar?

Nástroje k prozkoumání chybějících dat

Přehled chybějících dat

Manipulace chybějících dat

Vizualizace chybějících dat

Dříve se jmenoval ggmissing

The best thing to do with missing data is to not have any ***-- Gertrude Mary Cox***

Co balíček předpokládá?

Základní znalost v programu R

Zkušenost s vytvářením grafů v balíčku ggplot2

Zkušenost s balíčkem dplyr

Výhodou je mít i balíček visdat - v mnohém se doplňují

Jsou v datasetu nějaké chybějící hodnoty?

any_na() - odpověď TRUE - je tu alespoň 1 NA, FALSE - nenalezeno žádné NA

are_na() - o každé hodnotě řekne, jestli je NA nebo ne

n_miss() - počet NA

n_complete()

prop_miss() - poměr chybějících a nechybějících

miss_var_summary() - počet chybějících hodnot v každé proměnné a kolik je to procent

miss_case_summary() - počet a procenta chybějících každému subjektu

miss_var_span() - vypočítá počet chybějících hodnot v zadané proměnné pro opakující se rozpětí

Replacing missing values

ideal = NA

Může být ale kódováno špatně (např. “missing”, “not available”, “N/A”)

Dřív než tyto hodnoty nahradíme NA, měli bychom vědět, jak velký problém to je.

miss_scan_count(search = list(“N/A”)) - ukáže nám, v které proměnné se N/A nachází a kolikrát

replace_with_na(replace = list(*variable* = “N/A”)) - změní N/A na NA

je to zdlouhavé, musí se mnohokrát opakovat příkaz s touto funkcí, zjednodušení:

replace_with_na_all()

replace_with_na_at() - pouze na vybrané části přemění

replace_with_na_if() - přemění hodnoty v proměnných, které splňují nějakou podmínku (numeric, character)

Vizualizace chybějících hodnot

Klasicky ggplot2 nepracuje s missing values a z grafu je odstraní

vis_miss() - znázorní přehlednou vizualiaci missing value, kolik procent každé proměnné je missing a která pozorování to jsou

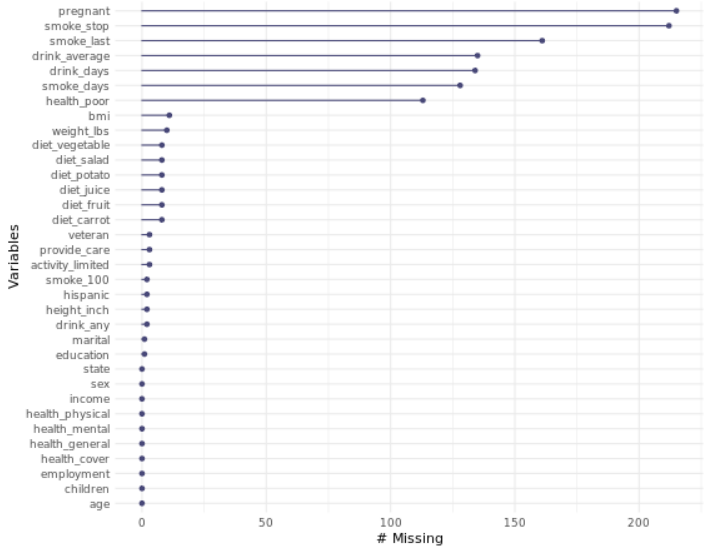
gg_miss_var(x, facet = y) - vizualizace missing value v proměnných, lze zvolit facet a znázornit tak missing ve všech hodnotách dané proměnné

gg_miss_case() - vizualizace missing value v pozorováních

gg_miss_upset() - celkový pohled na chybějící hodnoty

gg_miss_fct(x, fct = y)

gg_miss_span()



MCAR

To jsou missing value, která nemají žádnou pozorovanou spojitost s daty. Nedokážeme ani odhadnout, jaké hodnoty by měla missing value nabývat.

Doporučuje se odstranit pozorování s těmito missing value.

Zmenší se velikost vzorku, ale tyto data nebudou ovlivňovat analýzy.

MAR

Záleží na pozorovaných datech - např. pozorujeme, že s vysokými hodnotami některé proměnné, je v jiné proměnné NA

Smazat tyto NA není ideální

MNAR

Má souvislost v datech, ale nemůžeme ji pozorovat. Např. zároveň s chybějícíma hodnotama v jedné proměnné, se zároveň objevují chybějící hodnoty v jiné proměnné.

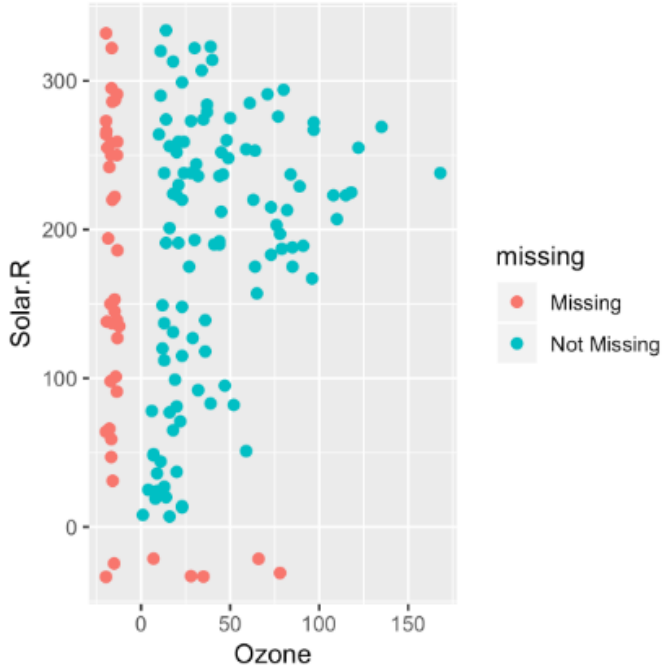
Data by byla zkreslena, kdybychom tyto missing value vymazali.

Visualizing missingness across two variables

ggplot(*dataframe*, aes(x = *variable*, y = *variable2*)) +

geom_miss_point()

- zaznamená všechny hodnoty, ale zvýrazní ty, které mají hodnoty jen z jedné proměnné



Impute missing values

Impute data a poté je znovu visualizujeme, abychom viděli rozdíl.

imput_below()

imput_below_if()

imput_below_at()

imput_below_all()

Střední hodnoty obecně nejsou dobré pro imputování missing values, doporučuje se nepoužívat.

Imputace se pak provádí dobře pomocí balíčku simputation.

Zdroje

https://cran.r-project.org/web/packages/naniar/vignettes/getting-started-w-naniar.html

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

https://www.rdocumentation.org/packages/naniar/versions/0.0.4.9000

https://www.datacamp.com/courses/dealing-with-missing-data-in-r