THE MAGIC OF REGEX

SARAH HIRSCH

WHAT IS REGEX?

- Short for “regular expression”
- System of codes used to generalize character strings
- Used in many programming languages, with small differences between versions and languages

Extremely useful for data cleaning and automation.

- Find and remove or replace systematic typos
- Perform fuzzy searches on data
- Conditionally loop through character strings
- Whatever else your coding heart desires!

SO… HOW DO I USE IT?

Common symbols:

\\d – digits

\\D – not digits

\\w – “word characters” (a-z, A-Z, 0-9, _)

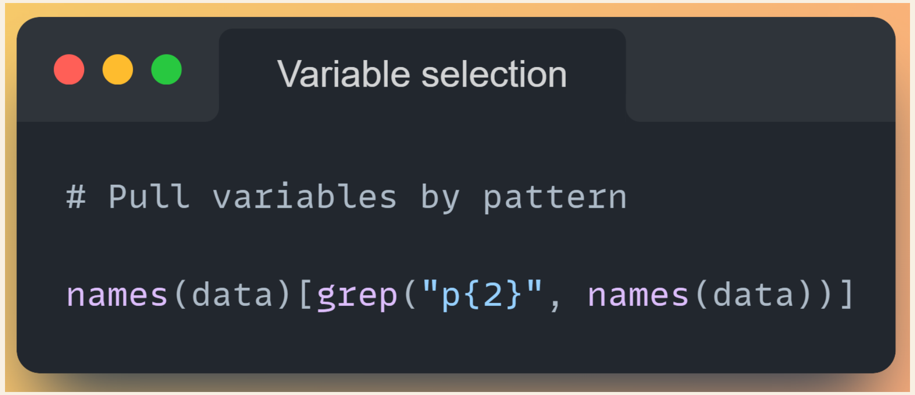
[[:alpha:]] – a-z or A-Z

[[:space:]] – blank space

+, *, {a}, {a,b} - one or more, 0 or more, *a* amount, *a* to *b* amount

^, $ - start of string, end of string

SOME COMMON USE CASES

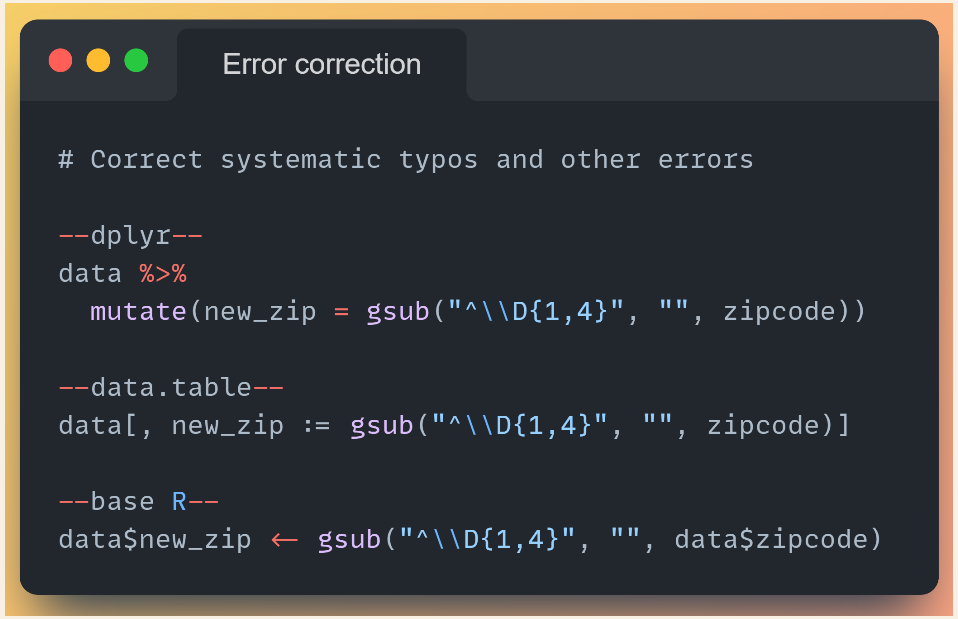


Variable selection

```
# Pull variables by pattern

names(data)[grep("p{2}", names(data))]
```

Please forgive the uneven cropping ☺



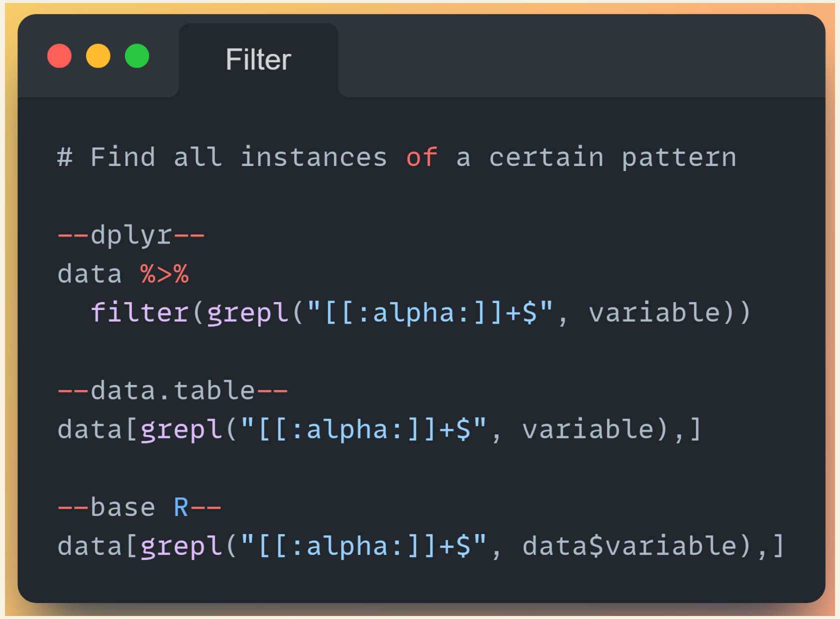
Error correction

```
# Correct systematic typos and other errors

--dplyr--
data %>%
  mutate(new_zip = gsub("^\\D{1,4}", "", zipcode))

--data.table--
data[, new_zip := gsub("^\\D{1,4}", "", zipcode)]

--base R--
data$new_zip ← gsub("^\\D{1,4}", "", data$zipcode)
```



Filter

```
# Find all instances of a certain pattern

--dplyr--
data %>%
  filter(grepl("[[:alpha:]]+$", variable))

--data.table--
data[grepl("[[:alpha:]]+$", variable),]

--base R--
data[grepl("[[:alpha:]]+$", data$variable),]
```