SOP for dataset Quality Control (QC)

Needed libraries: tidyverse, skimr, nanir, lubridate

1. Load necessary libraries:

library(tidyverse)

library(skimr)

library(naniar)

library(lubridate)

1. Load your file

File1 <- read.csv(file.choose(), stringsAsFactors=FALSE) #or

File2 <- read.delim("file2.txt") #or

File3 <- read.delim("file3.tsv", header = TRUE)

1. Inspect structure

glimpse(File)

skim(File)

1. Check missing value

File %>% summarise\_all(~sum(is.na(.)))

vis\_miss(File) # visualize missingness

1. Validate data types

sapply(File, class)

1. Check duplicates in unique IDs

sum(duplicated(File$Patient\_ID))

1. Identify outliers for numeric fields

File %>%

summarise(across(where(is.numeric),

list(min = min, max = max, mean = mean, sd = sd),

na.rm = TRUE))

1. Factor levels check

File %>% count(region)

unique(File$sex)

1. Date validation

file %>% summarise(min\_date = min(report\_date, na.rm = TRUE),

max\_date = max(report\_date, na.rm = TRUE))

1. Create a summary QC report

qc\_summary <- File %>%

summarise(

total\_rows = n(),

missing\_any = sum(!complete.cases(.)),

duplicates = sum(duplicated(Patient\_ID)),

mean\_age = mean(age, na.rm = TRUE),

invalid\_dates = sum(symptom\_onset > report\_date, na.rm = TRUE)

)

write\_csv(qc\_summary, "QC\_summary.csv")

##bonus

1. Check table header

head(File)

1. Check numbers of columns of the file

ncol(File)

1. Check number of rows of the file

nrow(File)

Example of NKA activity in ion-poor water condition

> NKA\_IP <- read.delim("/Users/Sherry/Desktop/Documents/in\_Canada/study/NKA/NKA\_for\_2017/NKA\_2017\_FF/NKA\_IP\_2017\_FF.csv", sep = "\t")

> head(NKA\_IP)

Region.Pop.NKAIP

1 coasr,FRP1,2.056943

2 coasr,FRP2,2.395011

3 coasr,FRP3,3.200272

4 coasr,FRP4,2.903362

5 coasr,FRP5,3.558925

6 coasr,FRP6,2.659396

> NKA\_IP <- read.csv("/Users/Sherry/Desktop/Documents/in\_Canada/study/NKA/NKA\_for\_2017/NKA\_2017\_FF/NKA\_IP\_2017\_FF.csv", sep = "\t")

> head(NKA\_IP)

Region.Pop.NKAIP

1 coasr,FRP1,2.056943

2 coasr,FRP2,2.395011

3 coasr,FRP3,3.200272

4 coasr,FRP4,2.903362

5 coasr,FRP5,3.558925

6 coasr,FRP6,2.659396

> ##Inspect file structure

> glimpse(NKA\_IP)

Rows: 70

Columns: 1

$ Region.Pop.NKAIP <chr> "coasr,FRP1,2.056943", "coasr,FRP2,2.395011", "coasr,FRP3,3.200272", "coasr,FRP…

> skim(NKA\_IP)

── Data Summary ────────────────────────

Values

Name NKA\_IP

Number of rows 70

Number of columns 1

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Column type frequency:

character 1

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Group variables None

── Variable type: character ──────────────────────────────────────────────────────────────────────────────

skim\_variable n\_missing complete\_rate min max empty n\_unique whitespace

1 Region.Pop.NKAIP 0 1 16 20 0 70 0

> NKA\_IP <- read.delim("/Users/Sherry/Desktop/Documents/in\_Canada/study/NKA/NKA\_for\_2017/NKA\_2017\_FF/NKA\_IP\_2017\_FF.csv", sep="\t") %>%

+ # Separate the combined first column into multiple columns

+ separate(col = 1, into = c("Region", "Pop", "NKAIP"),

+ sep = ",", convert = TRUE) %>%

+ # Clean up any whitespace

+ mutate(across(everything(), ~trimws(.)))

> glimpse(NKA\_IP)

Rows: 70

Columns: 3

$ Region <chr> "coasr", "coasr", "coasr", "coasr", "coasr", "coasr", "coasr", "coasr", "coasr", "coasr",…

$ Pop <chr> "FRP1", "FRP2", "FRP3", "FRP4", "FRP5", "FRP6", "FRP7", "FRP8", "CR1", "CR2", "CR3", "CR4…

$ NKAIP <chr> "2.056943", "2.395011", "3.200272", "2.903362", "3.558925", "2.659396", "3.218156", "2.16…

> ##Check missing value

> NKA\_IP %>% summarise\_all(~sum(is.na(.)))

Region Pop NKAIP

1 0 0 0

> vis\_miss(NKA\_IP) ## graph the result, as the result is 0, there in not bar or anything showed on the graph

> ##Validate data types

> sapply(NKA\_IP, class)

Region Pop NKAIP

"character" "character" "character"

> ##Check duplicates in unique IDs

> sum(duplicated(NKA\_IP$Pop))

[1] 0

> ##factor levels check

> NKA\_IP %>% count(Region)

Region n

1 coasl 20

2 coasr 25

3 inll 25

> unique(NKA\_IP$Region)

[1] "coasr" "coasl" "inll"

> # Identify outliers for numeric fields

> NKA\_IP %>%

+ summarise(across(where(is.numeric),

+ list(min = min, max = max, mean = mean, sd = sd),

+ na.rm = TRUE))

data frame with 0 columns and 1 row

Warning message:

There was 1 warning in `summarise()`.

ℹ In argument: `across(...)`.

Caused by warning:

! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.

Supply arguments directly to `.fns` through an anonymous function instead.

# Previously

across(a:b, mean, na.rm = TRUE)

# Now

across(a:b, \(x) mean(x, na.rm = TRUE))

This warning is displayed once every 8 hours.

Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was generated.

> # Identify outliers for numeric fields

> NKA\_IP <- NKA\_IP %>%

+ mutate(NKAIP = as.numeric(NKAIP)) ##NKA\_IP is character as indicated from above command

> NKA\_IP %>%

+ summarise(across(where(is.numeric),

+ list(min = min, max = max, mean = mean, sd = sd),

+ na.rm = TRUE))

NKAIP\_min NKAIP\_max NKAIP\_mean NKAIP\_sd

1 1.254103 6.80641 3.284215 1.343178

> # Date validation

> NKA\_IP %>% summarise(min\_date = min(NKAIP, na.rm = TRUE),

+ max\_date = max(NKAIP, na.rm = TRUE))

min\_date max\_date

1 1.254103 6.80641

> qc\_summary <- NKA\_IP %>%

+ summarise(

+ total\_rows = n(),

+ missing\_any = sum(!complete.cases(.)),

+ duplicates = sum(duplicated(Pop)),

+ mean\_NKAIP = mean(NKAIP, na.rm = TRUE),

+ )

> write\_csv(qc\_summary, "/Users/Sherry/Desktop/Documents/in\_Canada/study/NKA/NKA\_for\_2017/NKA\_2017\_FF/QC\_summary.csv")