Exam 2: Summary Sheet

Factors: Used to Represent Categorical Data.

`factor(x)` Converts a vector `x` to a factor. `class(x)` `levels`

`nlevels` `table(x)` `fct_reorder` `fct_collapse`

To: Statistical Modelling (lm(), glm()) treat factors as categorical predictors.

- Visualization: handle factors appropriately

Strings: Sequence of characters used to represent text data

- Creating Strings: `""` (")

Operations

Lenght `nchar("hello") = 5`

`tolower("HELLO") = hello`

`toupper("world") = WORLD`

`str_detect("banana", pattern = "ana") = TRUE`

`str_locate( // )` return start & end position

`str_extract( // )` extracts first number

`str_extractall`

1. Replacing Patterns (strings)

`str_replace("banana", pattern = "a", replacement = "o")` (replace the first match)

`str_replace_all("banana", pattern = "a", replacement = "o")` (replaces all matches)

Regular Expressions: for pattern matching in strings

- `.` any characters
- `*` zero or more
- `+` one or more
- `?` zero or none
- `[]` character set
- `()` grouping
- `\d` digit
- `\w` word character
- `\s` whitespace

DATA IMPORT

`read.csv("data.csv")`: To read a comma separated value

`header` = TRUE and FALSE; `sep` `","` to separate

`na.strings = c("", "NA", "-99")`

- `read_csv` = Faster than read.csv but needs the readr
- `fread`: For reading larger files fastly

To read other file types

- Excel: `read_excel(data)` from `readxl` package
- JSON: From `jsonlite` package
- Databases: Packages: Uses DBI packages like DBI

To inspect (EDA)

`head()`

`tail()` - some clues

`str` = structure of data frame

`summary` = Summary stats for every

Troublesbooting Import

- Incorrect data types check `str()` with as. functions and readr

`glimpse`: To inspect the structure in a more concise way

Don't forget to check for missing values like `na.strings`