Data Import

This document will show how to import data.

Import the FAS Litters CSV

```
litters_df = read_csv("data/FAS_litters.csv")
```

```
## Rows: 49 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): Group, Litter Number, GD0 weight, GD18 weight
## dbl (4): GD of Birth, Pups born alive, Pups dead @ birth, Pups survive
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
litters_df = janitor::clean_names(litters_df)
```

Look at the dataset

```
litters_df
```

```
## # A tibble: 49 x 8
##    group litter_number   gd0_weight gd18_weight gd_of_birth pups_born_alive
##    <chr> <chr>           <chr>      <chr>             <dbl>           <dbl>
##  1 Con7  #85             19.7       34.7                 20               3
##  2 Con7  #1/2/95/2       27         42                   19               8
##  3 Con7  #5/5/3/83/3-3   26         41.4                 19               6
##  4 Con7  #5/4/2/95/2     28.5       44.1                 19               5
##  5 Con7  #4/2/95/3-3     <NA>       <NA>                 20               6
##  6 Con7  #2/2/95/3-2     <NA>       <NA>                 20               6
##  7 Con7  #1/5/3/83/3-3/2 <NA>       <NA>                 20               9
##  8 Con8  #3/83/3-3       <NA>       <NA>                 20               9
##  9 Con8  #2/95/3         <NA>       <NA>                 20               8
## 10 Con8  #3/5/2/2/95     28.5       <NA>                 20               8
## # i 39 more rows
## # i 2 more variables: pups_dead_birth <dbl>, pups_survive <dbl>
```

```
head(litters_df)
```

```
## # A tibble: 6 x 8
##   group litter_number gd0_weight gd18_weight gd_of_birth pups_born_alive
##   <chr> <chr>         <chr>      <chr>             <dbl>           <dbl>
## 1 Con7  #85           19.7       34.7                 20               3
## 2 Con7  #1/2/95/2     27         42                   19               8
## 3 Con7  #5/5/3/83/3-3 26         41.4                 19               6
## 4 Con7  #5/4/2/95/2   28.5       44.1                 19               5
## 5 Con7  #4/2/95/3-3   <NA>       <NA>                 20               6
## 6 Con7  #2/2/95/3-2   <NA>       <NA>                 20               6
## # i 2 more variables: pups_dead_birth <dbl>, pups_survive <dbl>
```

```
tail(litters_df, 10)
```

```
## # A tibble: 10 x 8
##    group litter_number gd0_weight gd18_weight gd_of_birth pups_born_alive
##    <chr> <chr>         <chr>      <chr>             <dbl>           <dbl>
##  1 Mod8  #7/110/3-2    27.5       46                   19               8
##  2 Mod8  #2/95/2       28.5       44.5                 20               9
##  3 Mod8  #82/4         33.4       52.7                 20               8
##  4 Low8  #53           21.8       37.2                 20               8
##  5 Low8  #79           25.4       43.8                 19               8
##  6 Low8  #100          20         39.2                 20               8
##  7 Low8  #4/84         21.8       35.2                 20               4
##  8 Low8  #108          25.6       47.5                 20               8
##  9 Low8  #99           23.5       39                   20               6
## 10 Low8  #110          25.5       42.7                 20               7
## # i 2 more variables: pups_dead_birth <dbl>, pups_survive <dbl>
```

```
view(litters_df)
```

Learning Assessment

First load the FAS_pups.csv file using the relative path

```
pups_df = read_csv("data/FAS_pups.csv")
```

```
## Rows: 313 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Litter Number, PD ears
## dbl (4): Sex, PD eyes, PD pivot, PD walk
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
pups_df = janitor::clean_names(pups_df)

pups_df
```

```
## # A tibble: 313 x 6
##    litter_number   sex pd_ears pd_eyes pd_pivot pd_walk
##    <chr>         <dbl> <chr>     <dbl>    <dbl>   <dbl>
##  1 #85               1 4            13        7      11
##  2 #85               1 4            13        7      12
##  3 #1/2/95/2         1 5            13        7       9
##  4 #1/2/95/2         1 5            13        8      10
##  5 #5/5/3/83/3-3     1 5            13        8      10
##  6 #5/5/3/83/3-3     1 5            14        6       9
##  7 #5/4/2/95/2       1 .            14        5       9
##  8 #4/2/95/3-3       1 4            13        6       8
##  9 #4/2/95/3-3       1 4            13        7       9
## 10 #2/2/95/3-2       1 4            NA        8      10
## # i 303 more rows
```

Use absolute path.

```
pups_df = read_csv("~/Documents/School/Fall2024/BIST P8105/data_wrangling_I/data/FAS_pups.csv")
```

Look at read_csv options

col_names and skipping rows

```
litters_df =
  read_csv(
    file="data/FAS_litters.csv",
    col_names = FALSE,
  )
```

By default TRUE: 1st is colname
if FALSE, 1st is data

```
## Rows: 50 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (8): X1, X2, X3, X4, X5, X6, X7, X8
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
spec(litters_df)
```

```
## cols(
##   X1 = col_character(),
##   X2 = col_character(),
##   X3 = col_character(),
##   X4 = col_character(),
##   X5 = col_character(),
##   X6 = col_character(),
##   X7 = col_character(),
##   X8 = col_character()
## )
```

```
show_col_types = FALSE
```

What about missing data

```
litters_df =
  read_csv(
    file = "data/FAS_litters.csv",
    na = c("NA", "", ".")
  )
```

```
## Rows: 49 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): Group, Litter Number
## dbl (6): GD0 weight, GD18 weight, GD of Birth, Pups born alive, Pups dead @ ...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
litters_df = janitor::clean_names(litters_df)

pull(litters_df, gd0_weight)
```

```
##  [1] 19.7 27.0 26.0 28.5   NA   NA   NA   NA   NA 28.5 28.0   NA   NA   NA   NA
## [16] 17.0 21.4   NA   NA   NA 28.0 23.5 22.6   NA 21.7 24.4 19.5 24.3 22.6 22.2
## [31] 23.8 22.6 23.8 25.5 23.9 24.5   NA   NA 26.9 27.5 28.5 33.4 21.8 25.4 20.0
## [46] 21.8 25.6 23.5 25.5
```

What if we code `group` as a factor variable?

```
litters_df =
  read_csv(
    file = "data/FAS_litters.csv",
    na = c("NA", "", "."),
    col_types = cols(
      Group = col_factor()
    )
  )
```

Importing an excel file

Import MLB 2011 summary data

```
mlb_df = read_excel("data/mlb11.xlsx", sheet = "mlb11")
```

Import SAS data

```
pulse_df = read_sas("data/public_pulse_data.sas7bdat")
```

Never use read.csv()

```
litter_df = read_csv("data/FAS_litters.csv")
```

```
## Rows: 49 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (4): Group, Litter Number, GD0 weight, GD18 weight
## dbl (4): GD of Birth, Pups born alive, Pups dead @ birth, Pups survive
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

Never do this either:

```
litters_df$L
```

```
## Warning: Unknown or uninitialised column: ‘L‘.
```

```
## NULL
```