viz_and_eda

yc4384_Yangyang_Chen

2025-06-10

```
pulse_df =
  haven::read_sas("data/public_pulse_data.sas7bdat") |>
  janitor::clean_names()
```

```
pulse_tidy_df =
  pulse_df |>
  pivot_longer(bdi_score_bl:bdi_score_12m,
               names_to = "visit",
               names_prefix = "bdi_score_",
               values_to = "bdi") |>
  mutate(
    visit = replace(visit, visit == "bl", "00m"),
    visit = factor(visit)
  )
```

```
litters_wide =
  read_csv(
    "data/FAS_litters.csv",
    na = c("NA", ".", "")
  ) |>
  janitor::clean_names() |>
  select(litter_number, ends_with("weight")) |>
  pivot_longer(
    gd0_weight:gd18_weight,
    names_to = "gd",
    values_to = "weight"
  ) |>
  mutate(
    gd = case_match(
      gd,
      "gd0_weight" ~ 0,
      "gd18_weight" ~ 18
    )
  )
```

```
## 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.
```

```
analysis_result =
  tibble(
    group = c("treatment", "treatment", "placebo", "placebo"),
    time = c("pre", "post", "pre", "post"),
    mean = c(4, 8, 3.5, 4)
  )

analysis_result
```

```
## # A tibble: 4 x 3
##   group     time   mean
##   <chr>     <chr> <dbl>
## 1 treatment pre     4
## 2 treatment post    8
## 3 placebo   pre     3.5
## 4 placebo   post    4
```

```
pivot_wider(
  analysis_result,
  names_from = "time",
  values_from = "mean") |>
  knitr::kable()
```

group	pre	post
treatment	4.0	8
placebo	3.5	4

```
fellowship_ring =
  readxl::read_excel("./data/LotR_Words.xlsx", range = "B3:D6") |>
  mutate(movie = "fellowship_ring")

two_towers =
  readxl::read_excel("./data/LotR_Words.xlsx", range = "F3:H6") |>
  mutate(movie = "two_towers")

return_king =
  readxl::read_excel("./data/LotR_Words.xlsx", range = "J3:L6") |>
  mutate(movie = "return_king")
```

```
lotr_tidy =
  bind_rows(fellowship_ring, two_towers, return_king) |>
  janitor::clean_names() |>
  pivot_longer(
    female:male,
    names_to = "gender",
    values_to = "words") |>
  mutate(race = str_to_lower(race)) |>
  select(movie, everything())
```

```
lotr_tidy
```

```
## # A tibble: 18 x 4
##    movie           race   gender words
##    <chr>           <chr>  <chr>  <dbl>
##  1 fellowship_ring elf    female  1229
##  2 fellowship_ring elf    male     971
##  3 fellowship_ring hobbit female    14
##  4 fellowship_ring hobbit male    3644
##  5 fellowship_ring man    female     0
##  6 fellowship_ring man    male    1995
##  7 two_towers      elf    female   331
##  8 two_towers      elf    male     513
##  9 two_towers      hobbit female     0
## 10 two_towers      hobbit male    2463
## 11 two_towers      man    female   401
## 12 two_towers      man    male    3589
## 13 return_king     elf    female   183
## 14 return_king     elf    male     510
## 15 return_king     hobbit female     2
## 16 return_king     hobbit male    2673
## 17 return_king     man    female   268
## 18 return_king     man    male    2459
```

```
pup_df =
  read_csv(
    "./data/FAS_pups.csv",
    na = c("NA", "", ".")) |>
  janitor::clean_names() |>
  mutate(
    sex =
      case_match(
        sex,
        1 ~ "male",
        2 ~ "female"),
    sex = as.factor(sex))
```

```
## Rows: 313 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Litter Number
## dbl (5): Sex, PD ears, 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.
```

```
pup_df =
  read_csv(
    "data/FAS_pups.csv",
    na = c("NA", "", ".")
  ) |>
  janitor::clean_names() |>
  mutate(
    sex =
      case_match(
        sex,
        1 ~ "male",
        2 ~ "female"
      ),
    sex = as.factor(sex)
  )
```

```
## Rows: 313 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Litter Number
## dbl (5): Sex, PD ears, 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.
```

```
litter_df =
  read_csv(
    "./data/FAS_litters.csv",
    na = c("NA", ".", "")) |>
  janitor::clean_names() |>
  separate(group, into = c("dose", "day_of_tx"), sep = 3) |>
  relocate(litter_number) |>
  mutate(
    wt_gain = gd18_weight - gd0_weight,
    dose = str_to_lower(dose))
```

```
## 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.
```

```
surv_os =
  read_csv("data/surv_os.csv") |>
  janitor::clean_names() |>
  rename(id = what_is_your_uni, os = what_operating_system_do_you_use)
```

```
## Rows: 173 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): What is your UNI?, What operating system do you use?
##
## 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.
```

```r
surv_pr_git =
  read_csv("data/surv_program_git.csv") |>
  janitor::clean_names() |>
  rename(
    id = what_is_your_uni,
    prog = what_is_your_degree_program,
    git_exp = which_most_accurately_describes_your_experience_with_git)
```

```
## Rows: 135 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): What is your UNI?, What is your degree program?, Which most accurat...
##
## 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.
```

```r
left_join(surv_os, surv_pr_git)
```

```
## Joining with ‘by = join_by(id)‘
```

```
## Warning in left_join(surv_os, surv_pr_git): Detected an unexpected many-to-many relationship between
## i Row 7 of ‘x‘ matches multiple rows in ‘y‘.
## i Row 66 of ‘y‘ matches multiple rows in ‘x‘.
## i If a many-to-many relationship is expected, set ‘relationship =
##   "many-to-many"‘ to silence this warning.
```

```
## # A tibble: 175 x 4
##    id          os         prog  git_exp
##    <chr>       <chr>      <chr> <chr>
##  1 student_87  <NA>       MS    Pretty smooth: needed some work to connect Git,~
##  2 student_106 Windows 10 Other Pretty smooth: needed some work to connect Git,~
##  3 student_66  Mac OS X   MPH   Smooth: installation and connection with GitHub~
##  4 student_93  Windows 10 MS    Smooth: installation and connection with GitHub~
##  5 student_99  Mac OS X   MS    Smooth: installation and connection with GitHub~
##  6 student_115 Mac OS X   MS    Smooth: installation and connection with GitHub~
##  7 student_15  Windows 10 MPH   Pretty smooth: needed some work to connect Git,~
##  8 student_15  Windows 10 MPH   Pretty smooth: needed some work to connect Git,~
##  9 student_21  Windows 10 MPH   Pretty smooth: needed some work to connect Git,~
## 10 student_86  Mac OS X   <NA>  <NA>
## # i 165 more rows
```

```r
inner_join(surv_os, surv_pr_git)
```

```
## Joining with ‘by = join_by(id)‘
```

```
## Warning in inner_join(surv_os, surv_pr_git): Detected an unexpected many-to-many relationship betwee
## i Row 7 of ‘x‘ matches multiple rows in ‘y‘.
## i Row 66 of ‘y‘ matches multiple rows in ‘x‘.
## i If a many-to-many relationship is expected, set ‘relationship =
##   "many-to-many"‘ to silence this warning.
## # A tibble: 129 x 4
##    id          os         prog  git_exp
##    <chr>       <chr>      <chr> <chr>
##  1 student_87  <NA>       MS    Pretty smooth: needed some work to connect Git,~
##  2 student_106 Windows 10 Other Pretty smooth: needed some work to connect Git,~
##  3 student_66  Mac OS X   MPH   Smooth: installation and connection with GitHub~
##  4 student_93  Windows 10 MS    Smooth: installation and connection with GitHub~
##  5 student_99  Mac OS X   MS    Smooth: installation and connection with GitHub~
##  6 student_115 Mac OS X   MS    Smooth: installation and connection with GitHub~
##  7 student_15  Windows 10 MPH   Pretty smooth: needed some work to connect Git,~
##  8 student_15  Windows 10 MPH   Pretty smooth: needed some work to connect Git,~
##  9 student_21  Windows 10 MPH   Pretty smooth: needed some work to connect Git,~
## 10 student_59  Windows 10 MPH   Smooth: installation and connection with GitHub~
## # i 119 more rows
```

```
anti_join(surv_os, surv_pr_git)
```

```
## Joining with ‘by = join_by(id)‘
```

```
## # A tibble: 46 x 2
##    id          os
##    <chr>       <chr>
##  1 student_86  Mac OS X
##  2 student_91  Windows 10
##  3 student_24  Mac OS X
##  4 student_103 Mac OS X
##  5 student_163 Mac OS X
##  6 student_68  Other (Linux, Windows, 95, TI-89+, etc)
##  7 student_158 Mac OS X
##  8 student_19  Windows 10
##  9 student_43  Mac OS X
## 10 student_78  Mac OS X
## # i 36 more rows
```

```
anti_join(surv_pr_git, surv_os)
```

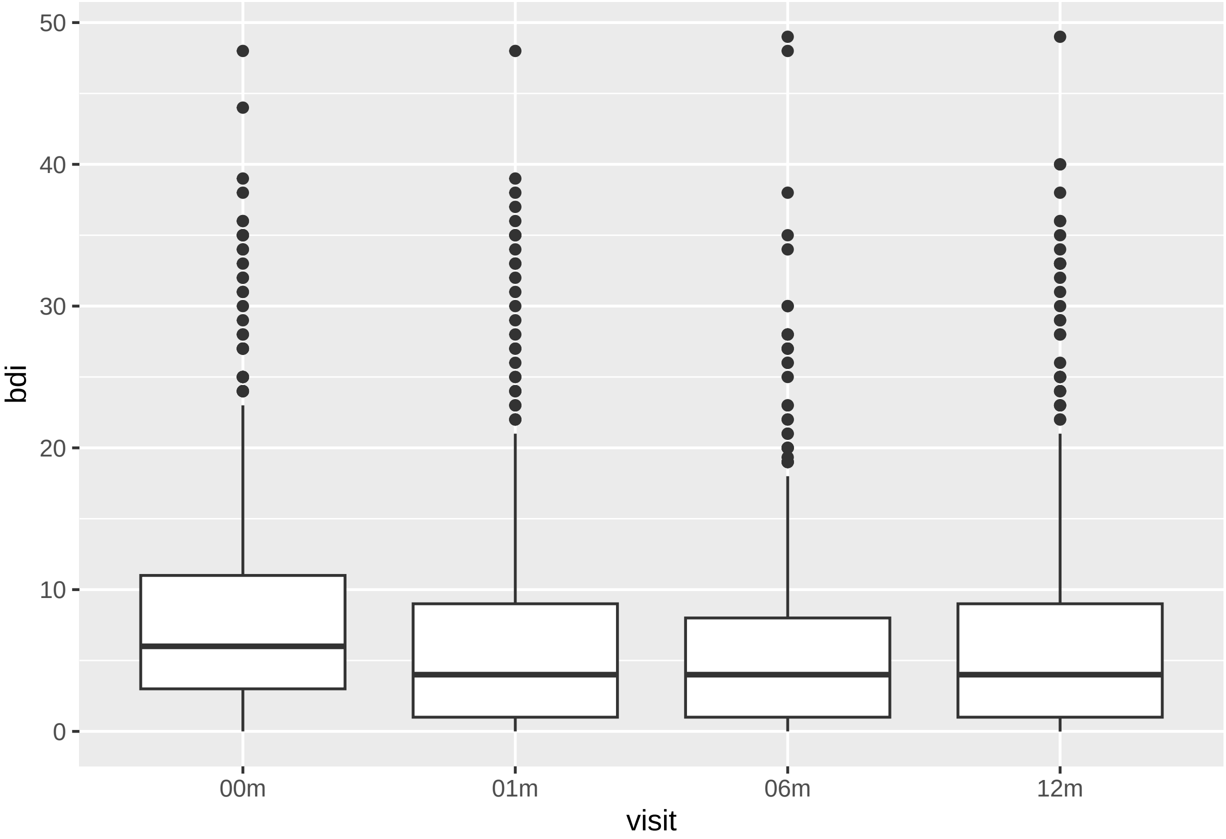
```
## Joining with ‘by = join_by(id)‘
```

```
## # A tibble: 15 x 3
##    id         prog  git_exp
##    <chr>      <chr> <chr>
##  1 <NA>       MPH   "Pretty smooth: needed some work to connect Git, GitHub, an~
##  2 student_17 PhD   "Pretty smooth: needed some work to connect Git, GitHub, an~
##  3 <NA>       MPH   "Pretty smooth: needed some work to connect Git, GitHub, an~
##  4 <NA>       MPH   "Pretty smooth: needed some work to connect Git, GitHub, an~
##  5 <NA>       MS    "Pretty smooth: needed some work to connect Git, GitHub, an~
##  6 student_53 MS    "Pretty smooth: needed some work to connect Git, GitHub, an~
##  7 <NA>       MS    "Smooth: installation and connection with GitHub was easy"
##  8 student_80 PhD   "Pretty smooth: needed some work to connect Git, GitHub, an~
##  9 student_16 MPH   "Smooth: installation and connection with GitHub was easy"
## 10 student_98 MS    "Smooth: installation and connection with GitHub was easy"
## 11 <NA>       MS    "Pretty smooth: needed some work to connect Git, GitHub, an~
## 12 <NA>       MS    "What’s \"Git\" ...?"
## 13 <NA>       MS    "Smooth: installation and connection with GitHub was easy"
## 14 <NA>       MPH   "Pretty smooth: needed some work to connect Git, GitHub, an~
## 15 <NA>       MS    "Pretty smooth: needed some work to connect Git, GitHub, an~
```

```
pulse_data =
  haven::read_sas("./data/public_pulse_data.sas7bdat") |>
  janitor::clean_names() |>
  pivot_longer(
    bdi_score_bl:bdi_score_12m,
    names_to = "visit",
    names_prefix = "bdi_score_",
    values_to = "bdi") |>
  select(id, visit, everything()) |>
  mutate(
    visit = recode(visit, "bl" = "00m"),
    visit = factor(visit, levels = str_c(c("00", "01", "06", "12"), "m"))) |>
  arrange(id, visit)


ggplot(pulse_data, aes(x = visit, y = bdi)) +
  geom_boxplot()
```

```
## Warning: Removed 879 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```


bdi
00m
01m
06m
12m
visit