Data Wrangling in R

Howard Nguyen

2020-01-02

LOAD PACKAGES

```
# Load packages
library(tidyverse)  # Loads the `tidyverse` collection
library(readxl)     # Reads CSV and Excel files
```

LOAD DATA

```
# Also convert several adjacent variables to factors
df <- read_csv("../data/state_trends.csv") |>
  select(state, region, psych_region, data_analysis) |>
  mutate(across(c(region:psych_region), as_factor)) |>
  print()
```

```
## Rows: 48 Columns: 34
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): state, state_code, region, psych_region, psy_reg, has_nba, has_nfl...
## dbl (23): population, sq_miles, pop_density, extraversion, agreeableness, co...
##
## 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.
```

```
## # A tibble: 48 x 4
##    state       region    psych_region                  data_analysis
##    <chr>       <fct>     <fct>                                 <dbl>
##  1 Alabama     South     Friendly and Conventional                35
##  2 Arizona     West      Relaxed and Creative                     35
##  3 Arkansas    South     Friendly and Conventional                40
##  4 California  West      Relaxed and Creative                     46
##  5 Colorado    West      Friendly and Conventional                35
##  6 Connecticut Northeast Temperamental and Uninhibited            40
##  7 Delaware    South     Temperamental and Uninhibited            43
##  8 Florida     South     Friendly and Conventional                35
##  9 Georgia     South     Friendly and Conventional                38
## 10 Idaho       West      Relaxed and Creative                     33
## # ... with 38 more rows
```

FILTER BY ONE VARIABLE

```
# "data_analysis" is a numeric variable
df |>
  filter(data_analysis > 50) |>
  arrange(desc(data_analysis)) |>  # Sorts output
  print()
```

```
## # A tibble: 4 x 4
##   state         region    psych_region                   data_analysis
##   <chr>         <fct>     <fct>                                  <dbl>
## 1 Maryland      South     Temperamental and Uninhibited             64
## 2 New York      Northeast Temperamental and Uninhibited             63
## 3 Massachusetts Northeast Temperamental and Uninhibited             62
## 4 Virginia      South     Relaxed and Creative                      56
```

```
# "psych_region" is a text variable
df |>
  filter(psych_region == "Relaxed and Creative") |>
  arrange(desc(data_analysis)) |>  # Sorts output
  print()
```

```
## # A tibble: 10 x 4
##    state          region psych_region         data_analysis
##    <chr>          <fct>  <fct>                        <dbl>
##  1 Virginia       South  Relaxed and Creative            56
##  2 California     West   Relaxed and Creative            46
##  3 Washington     West   Relaxed and Creative            41
##  4 North Carolina South  Relaxed and Creative            40
##  5 Utah           West   Relaxed and Creative            38
##  6 Arizona        West   Relaxed and Creative            35
##  7 Idaho          West   Relaxed and Creative            33
##  8 New Mexico     West   Relaxed and Creative            33
##  9 Oregon         West   Relaxed and Creative            31
## 10 Nevada         West   Relaxed and Creative            27
```

FILTER BY MULTIPLE VARIABLES

```
# "or" is the vertical pipe |
df |>
  filter(region == "South" |
    psych_region == "Relaxed and Creative") |>
  arrange(region, psych_region) |>  # Sorts output
  print(n = Inf)  # Print all rows
```

```
## # A tibble: 24 x 4
##    state          region psych_region                  data_analysis
##    <chr>          <fct>  <fct>                                 <dbl>
```

```
##  1 Alabama        South  Friendly and Conventional               35
##  2 Arkansas       South  Friendly and Conventional               40
##  3 Florida        South  Friendly and Conventional               35
##  4 Georgia        South  Friendly and Conventional               38
##  5 Kentucky       South  Friendly and Conventional               31
##  6 Louisiana      South  Friendly and Conventional               29
##  7 Mississippi    South  Friendly and Conventional               33
##  8 Oklahoma       South  Friendly and Conventional               29
##  9 South Carolina South  Friendly and Conventional               32
## 10 Tennessee      South  Friendly and Conventional               28
## 11 North Carolina South  Relaxed and Creative                    40
## 12 Virginia       South  Relaxed and Creative                    56
## 13 Delaware       South  Temperamental and Uninhibited           43
## 14 Maryland       South  Temperamental and Uninhibited           64
## 15 Texas          South  Temperamental and Uninhibited           40
## 16 West Virginia  South  Temperamental and Uninhibited           30
## 17 Arizona        West   Relaxed and Creative                    35
## 18 California     West   Relaxed and Creative                    46
## 19 Idaho          West   Relaxed and Creative                    33
## 20 Nevada         West   Relaxed and Creative                    27
## 21 New Mexico     West   Relaxed and Creative                    33
## 22 Oregon         West   Relaxed and Creative                    31
## 23 Utah           West   Relaxed and Creative                    38
## 24 Washington     West   Relaxed and Creative                    41
```

```
# "and" is the ampersand &
df |>
  filter(region == "South" &
    psych_region == "Relaxed and Creative") |>
  print()
```

```
## # A tibble: 2 x 4
##   state          region psych_region         data_analysis
##   <chr>          <fct>  <fct>                        <dbl>
## 1 North Carolina South  Relaxed and Creative            40
## 2 Virginia       South  Relaxed and Creative            56
```

```
# "not" is the exclamation point !
df |>
  filter(region == "South" &
    !psych_region == "Relaxed and Creative") |>
  arrange(psych_region, desc(data_analysis)) |>
  print()
```

```
## # A tibble: 14 x 4
##    state          region psych_region                 data_analysis
##    <chr>          <fct>  <fct>                                <dbl>
##  1 Arkansas       South  Friendly and Conventional               40
##  2 Georgia        South  Friendly and Conventional               38
##  3 Alabama        South  Friendly and Conventional               35
##  4 Florida        South  Friendly and Conventional               35
##  5 Mississippi    South  Friendly and Conventional               33
##  6 South Carolina South  Friendly and Conventional               32
```

```
##  7 Kentucky       South  Friendly and Conventional               31
##  8 Louisiana      South  Friendly and Conventional               29
##  9 Oklahoma       South  Friendly and Conventional               29
## 10 Tennessee      South  Friendly and Conventional               28
## 11 Maryland       South  Temperamental and Uninhibited           64
## 12 Delaware       South  Temperamental and Uninhibited           43
## 13 Texas          South  Temperamental and Uninhibited           40
## 14 West Virginia  South  Temperamental and Uninhibited           30
```

RECORDING

LOAD DATA

```
# Also convert all character variables to factors
df <- read_csv("../data/state_trends.csv") |>
  mutate(across(where(is_character), as_factor)) |>
  print()
```

```
## Rows: 48 Columns: 34
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (11): state, state_code, region, psych_region, psy_reg, has_nba, has_nfl...
## dbl (23): population, sq_miles, pop_density, extraversion, agreeableness, co...
##
## 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.
```

```
## # A tibble: 48 x 34
##    state  state~1 popul~2 sq_mi~3 pop_d~4 region psych~5 psy_reg extra~6 agree~7
##    <fct>  <fct>     <dbl>   <dbl>   <dbl> <fct>  <fct>   <fct>     <dbl>   <dbl>
##  1 Alaba~ AL       5.02e6   52420      96 South  Friend~ Friend~    55.5    52.7
##  2 Arizo~ AZ       7.15e6  113990      63 West   Relaxe~ Creati~    50.6    46.6
##  3 Arkan~ AR       3.01e6   53179      57 South  Friend~ Friend~    49.9    52.7
##  4 Calif~ CA       3.95e7  163695     242 West   Relaxe~ Creati~    51.4    49
##  5 Color~ CO       5.77e6  104094      55 West   Friend~ Friend~    45.3    47.5
##  6 Conne~ CT       3.61e6    5543     650 North~ Temper~ Uninhi~    57.6    38.6
##  7 Delaw~ DE       9.90e5    2489     398 South  Temper~ Uninhi~    47      38.8
##  8 Flori~ FL       2.15e7   65758     328 South  Friend~ Friend~    60.9    50.7
##  9 Georg~ GA       1.07e7   59425     180 South  Friend~ Friend~    63.2    60
## 10 Idaho  ID       1.84e6   83569      22 West   Relaxe~ Creati~    40.7    52.9
## # ... with 38 more rows, 24 more variables: conscientiousness <dbl>,
## #   neuroticism <dbl>, openness <dbl>, data_science <dbl>,
## #   artificial_intelligence <dbl>, machine_learning <dbl>, data_analysis <dbl>,
## #   business_intelligence <dbl>, spreadsheet <dbl>, statistics <dbl>,
## #   art <dbl>, dance <dbl>, museum <dbl>, basketball <dbl>, football <dbl>,
## #   baseball <dbl>, soccer <dbl>, hockey <dbl>, has_nba <fct>, has_nfl <fct>,
## #   has_mlb <fct>, has_mls <fct>, has_nhl <fct>, has_any <fct>, and ...
```

COMBINE CATEGORIES WITH RECODE

```
df |>
  mutate(relaxed = recode(psych_region,
    "Relaxed and Creative" = "yes",
    "Friendly and Conventional" = "no",
    .default = "no")) |>  # Sets default value
  select(state_code, psych_region, relaxed)
```

```
## # A tibble: 48 x 3
##    state_code psych_region                  relaxed
##    <fct>      <fct>                         <fct>
##  1 AL         Friendly and Conventional     no
##  2 AZ         Relaxed and Creative          yes
##  3 AR         Friendly and Conventional     no
##  4 CA         Relaxed and Creative          yes
##  5 CO         Friendly and Conventional     no
##  6 CT         Temperamental and Uninhibited no
##  7 DE         Temperamental and Uninhibited no
##  8 FL         Friendly and Conventional     no
##  9 GA         Friendly and Conventional     no
## 10 ID         Relaxed and Creative          yes
## # ... with 38 more rows
```

CREATE CATEGORIES WITH CASE_WHEN

?case_when # Help on `case_when`

```
df |>
  mutate(
    like_arts = case_when(
      art > 75 | dance > 75 | museum > 75 ~ "yes",
      TRUE ~ "no"  # All other values
    )
  ) |>
  select(state_code, like_arts, art:museum) |>
  arrange(desc(like_arts)) |>  # Put yes at top
  print(n = Inf)               # Show all cases
```

```
## # A tibble: 48 x 5
##    state_code like_arts   art dance museum
##    <fct>      <chr>     <dbl> <dbl>  <dbl>
##  1 AZ         yes          78    69     26
##  2 CA         yes          84    70     25
##  3 CO         yes          85    78     29
##  4 CT         yes          80    74     31
##  5 FL         yes          77    69     24
##  6 ID         yes          82    77     17
##  7 IL         yes          74    78     32
##  8 KS         yes          80    70     26
```

```
##  9 ME         yes          85    70     34
## 10 MD         yes          76    73     33
## 11 MA         yes          78    76     40
## 12 MI         yes          89    70     21
## 13 MN         yes          78    77     23
## 14 MO         yes          80    75     28
## 15 MT         yes          78    59     26
## 16 NH         yes          75    81     27
## 17 NM         yes          86    65     33
## 18 NY         yes          80    69     41
## 19 OR         yes         100    75     23
## 20 RI         yes          78    74     33
## 21 UT         yes          89   100     23
## 22 VT         yes          92    79     36
## 23 WA         yes          85    66     28
## 24 WI         yes          81    72     28
## 25 WY         yes          78    65     29
## 26 AL         no           65    65     19
## 27 AR         no           72    61     20
## 28 DE         no           72    70     26
## 29 GA         no           71    69     21
## 30 IN         no           74    68     26
## 31 IA         no           69    74     21
## 32 KY         no           71    65     24
## 33 LA         no           68    73     27
## 34 MS         no           68    63     21
## 35 NE         no           70    74     27
## 36 NV         no           74    68     33
## 37 NJ         no           74    74     24
## 38 NC         no           73    71     25
## 39 ND         no           73    66     14
## 40 OH         no           75    68     25
## 41 OK         no           72    64     26
## 42 PA         no           72    69     27
## 43 SC         no           72    69     20
## 44 SD         no           70    67     23
## 45 TN         no           67    65     26
## 46 TX         no           74    67     21
## 47 VA         no           74    74     34
## 48 WV         no           66    65     16
```

NEW VARIABLES

```
# LOAD PACKAGES ################################################
# RStudio will prompt you to download any packages that
# aren't already installed.

library(tidyverse)
```

CREATE DATA

```
# Create a small dataset with 1-7 data and a missing value
df_new <- tibble(
  x = 1:5,
  y = 7:3,
  z = c(2, 4, 3, 7, NA)
) |>
print()
```

```
## # A tibble: 5 x 3
##       x     y     z
##   <int> <int> <dbl>
## 1     1     7     2
## 2     2     6     4
## 3     3     5     3
## 4     4     4     7
## 5     5     3    NA
```

AVERAGE ACROSS VARIABLES

```
# Average variables with `rowMeans`
df_new %>% mutate(
  mean_xy  = rowMeans(across(x:y)),
  mean_xyz = rowMeans(across(x:z)),
  mean_xz  = rowMeans(across(c(x, z)))
)
```

```
## # A tibble: 5 x 6
##       x     y     z mean_xy mean_xyz mean_xz
##   <int> <int> <dbl>   <dbl>    <dbl>   <dbl>
## 1     1     7     2       4     3.33     1.5
## 2     2     6     4       4     4        3
## 3     3     5     3       4     3.67     3
## 4     4     4     7       4     5        5.5
## 5     5     3    NA       4    NA       NA
```

```
# Remove missing values by adding `na.rm = T`
df_new %>% mutate(
  mean_xy  = rowMeans(across(x:y), na.rm = T),
  mean_xyz = rowMeans(across(x:z), na.rm = T),
  mean_xz  = rowMeans(across(c(x, z)), na.rm = T)
)
```

```
## # A tibble: 5 x 6
##       x     y     z mean_xy mean_xyz mean_xz
##   <int> <int> <dbl>   <dbl>    <dbl>   <dbl>
## 1     1     7     2       4     3.33     1.5
## 2     2     6     4       4     4        3
```

```
## 3     3     5     3       4     3.67     3
## 4     4     4     7       4     5        5.5
## 5     5     3    NA       4     4        5
```

REVERSE CODING

```
df_new %>%
  mutate(y_r = 8 - y) |>  # Create reversed variable
  select(x, y_r, z) |>    # Select and reorder variables
  mutate(                 # Compute average scores
    mean_xy  = rowMeans(across(c(x, y_r)), na.rm = T),
    mean_xyz = rowMeans(across(c(x, y_r, z)), na.rm = T),
    mean_xz  = rowMeans(across(c(x, z)), na.rm = T)
  )
```

```
## # A tibble: 5 x 6
##       x   y_r     z mean_xy mean_xyz mean_xz
##   <int> <dbl> <dbl>   <dbl>    <dbl>   <dbl>
## 1     1     1     2       1     1.33     1.5
## 2     2     2     4       2     2.67     3
## 3     3     3     3       3     3        3
## 4     4     4     7       4     5        5.5
## 5     5     5    NA       5     5        5
```

```
# For a 1-n scale, use (n + 1) - x
# So, for a 1-7 scale, use 8 - x
# So, for a 1-10 scale, use 11 - x
#
# For a 0-n scale, use n - x
# So, for a 0-5 scale, use 5 - x
# So, for a 0-10 scale, use 10 - x
#
# For a -n to +n scale, use x * -1
# So, for a -3 to +3 scale, use x * -1
# So, for a -10 to +10 scale, use x * -1
```