AE_06

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(scales)
```

```

Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```
library(ggthemes)
library(glue)

ggplot2::theme_set(ggplot2::theme_minimal(base_size = 14))
```

```
options(width = 65)
```

```
hotels <- read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data
```

```
Rows: 119390 Columns: 32
-- Column specification --------------------------------------
Delimiter: ","
chr  (13): hotel, arrival_date_month, meal, country, market_s...
dbl  (18): is_canceled, lead_time, arrival_date_year, arrival...
date  (1): reservation_status_date

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.
```

The problem that this rowwise operation has is that any value that is =0 or NA will return the total guests as NA which is not what we want.

```
hotels |>
  select(adults, children, babies) |>
  mutate(guests = sum(c(adults, children, babies)))
```

```
# A tibble: 119,390 x 4
   adults children babies guests
    <dbl>    <dbl>  <dbl>  <dbl>
 1      2        0      0     NA
 2      2        0      0     NA
 3      1        0      0     NA
 4      1        0      0     NA
 5      2        0      0     NA
 6      2        0      0     NA
 7      2        0      0     NA
 8      2        0      0     NA
 9      2        0      0     NA
10      2        0      0     NA
# i 119,380 more rows
```

Making sure that each adults children and babies are greater than zero will allow the total guests column to work. using rowwise() will help understand and make sure you get the right outcome for total guests.

```r
hotels |>
  select(adults, children, babies) |>
  rowwise() |>
  mutate(guests = sum(c(adults, children, babies))) |>
  filter(adults > 0, children > 0, babies > 0)
```

```
# A tibble: 172 x 4
# Rowwise:
   adults children babies guests
    <dbl>    <dbl>  <dbl>  <dbl>
 1      2        1      1      4
 2      2        1      1      4
 3      2        1      1      4
 4      2        1      1      4
 5      2        1      1      4
 6      2        1      1      4
 7      2        1      1      4
 8      2        2      1      5
 9      2        2      1      5
10      1        2      1      4
# i 162 more rows
```

```r
hotels |>
  summarise(across(.cols = starts_with("stays"), mean)) |>
  glimpse()
```

```
Rows: 1
Columns: 2
$ stays_in_weekend_nights <dbl> 0.9275986
$ stays_in_week_nights    <dbl> 2.500302
```

```r
hotels |>
  group_by(hotel, is_canceled) |>
  summarise(
    across(.cols = starts_with("stays"), list(mean = mean, sd = sd), .names = "{.fn}_{.col}")
  ) |>
  glimpse()
```

```
`summarise()` has grouped output by 'hotel'. You can override
using the `.groups` argument.
```

```
Rows: 4
Columns: 6
Groups: hotel [2]
$ hotel                         <chr> "City Hotel", "City Hotel"~
$ is_canceled                   <dbl> 0, 1, 0, 1
$ mean_stays_in_weekend_nights <dbl> 0.8006836, 0.7875053, 1.13~
$ sd_stays_in_weekend_nights   <dbl> 0.8615080, 0.9168195, 1.14~
$ mean_stays_in_week_nights    <dbl> 2.122934, 2.266781, 3.0089~
$ sd_stays_in_week_nights      <dbl> 1.400799, 1.526787, 2.4507~
```

```
hotels |>
  group_by(hotel, is_canceled) |>
  summarise(
    across(.cols = starts_with("stays"), list(mean = mean, sd = sd), .names = "{.fn}_{.col}"
    .groups = "drop"
  )
```

```
# A tibble: 4 x 6
  hotel is_canceled mean_stays_in_weeken~1 sd_stays_in_weekend_~2
  <chr>       <dbl>                  <dbl>                  <dbl>
1 City~           0                  0.801                  0.862
2 City~           1                  0.788                  0.917
3 Reso~           0                  1.13                   1.14
4 Reso~           1                  1.34                   1.14
# i abbreviated names: 1: mean_stays_in_weekend_nights,
#   2: sd_stays_in_weekend_nights
# i 2 more variables: mean_stays_in_week_nights <dbl>,
#   sd_stays_in_week_nights <dbl>
```

This gives mean number of stays giving is_canceled and which hotel/resort.

```
hotels_summary <- hotels |>
  group_by(hotel, is_canceled) |>
  summarise(
    across(
      .cols = starts_with("stays"),
      list(mean = mean),
      .names = "{.fn}_{.col}"
    ),
    .groups = "drop"
  )
```

```
hotels_summary
```

```
# A tibble: 4 x 4
  hotel is_canceled mean_stays_in_weeken~1 mean_stays_in_week_n~2
  <chr>       <dbl>                  <dbl>                  <dbl>
1 City~           0                  0.801                   2.12
2 City~           1                  0.788                   2.27
3 Reso~           0                  1.13                    3.01
4 Reso~           1                  1.34                    3.44
# i abbreviated names: 1: mean_stays_in_weekend_nights,
#   2: mean_stays_in_week_nights
```

```
hotels_summary %>%
  pivot_longer(c(3,4), names_to = "stay_type", values_to = "mean_rate") %>%
  mutate(stay_type = if_else(str_detect(stay_type, "weekend"), "Weekend", "Weekday")) %>%
  mutate(is_canceled = recode(is_canceled, "0" = "Not Canceled", "1" = "Canceled")) %>%
  arrange(hotel, is_canceled) %>%
  ggplot(aes(x = is_canceled, y = mean_rate, linetype = hotel)) +
  geom_point(aes(color = hotel)) +
  facet_wrap(~ stay_type) +
  geom_line(aes(group = hotel, color = hotel))
```