Wrangling practice: one table

DATA 202 21FA, based on datasciencebox.org

Q&A

> When are `ggplot2` vs `dplyr` used in the real world?

Usually they're used *together*: wrangle-wrangle-wrangle-plot. Reports often include both *tables* and *graphics*.

> Are `mutate`, `select`, etc. universal?

These are only `tidyverse`, but SQL has very similar concepts.

> How to connect and join tables?

Next week! (You're welcome to read ahead.)

> Can you pipe too much?

No intrinsic limit, but: split into functions, simplify your analysis.

Today: Practice thinking about grammar of transformation

Friday: in-lab practice

Data: Hotel bookings

- Data from two hotels: one resort and one city hotel
- Observations: Each row represents a hotel booking
- You can try it: Application Exercise

```
hotels <- read_csv("data/hotels.csv")
head(hotels)
```

```
## # A tibble: 6 × 32
##   hotel        is_canceled lead_time arrival_date_ye… arrival_date_m
##   <chr>              <dbl>     <dbl>            <dbl> <chr>
## 1 Resort Hotel           0       342             2015 July
## 2 Resort Hotel           0       737             2015 July
## 3 Resort Hotel           0         7             2015 July
## 4 Resort Hotel           0        13             2015 July
## 5 Resort Hotel           0        14             2015 July
## 6 Resort Hotel           0        14             2015 July
## # … with 27 more variables: arrival_date_week_number <dbl>,
## #   arrival_date_day_of_month <dbl>,
## #   stays_in_weekend_nights <dbl>, stays_in_week_nights <dbl>,
## #   adults <dbl>, children <dbl>, babies <dbl>, meal <chr>,
## #   country <chr>, market_segment <chr>,
```

Source: TidyTuesday

ascending / descending order

```
hotels %>%
  select(adults, children, bab
  arrange(babies)
```

```
## # A tibble: 119,390 × 3
##   adults children babies
##    <dbl>    <dbl>  <dbl>
## 1      2        0      0
## 2      2        0      0
## 3      1        0      0
## 4      1        0      0
## 5      2        0      0
## 6      2        0      0
## # … with 119,384 more rows
```

ascending / descending order

```
hotels %>%
  select(adults, children, bab
  arrange(babies)
```

```
## # A tibble: 119,390 × 3
##   adults children babies
##    <dbl>    <dbl>  <dbl>
## 1      2        0      0
## 2      2        0      0
## 3      1        0      0
## 4      1        0      0
## 5      2        0      0
## 6      2        0      0
## # … with 119,384 more rows
```

```
hotels %>%
  select(adults, children, bab
  arrange(desc(babies))
```

```
## # A tibble: 119,390 × 3
##   adults children babies
##    <dbl>    <dbl>  <dbl>
## 1      2        0     10
## 2      1        0      9
## 3      2        0      2
## 4      2        0      2
## 5      2        0      2
## 6      2        0      2
## # … with 119,384 more rows
```

ascending / descending order

```
hotels %>%
  select(adults, children, bab
  arrange(babies)
```

```
## # A tibble: 119,390 × 3
##   adults children babies
##    <dbl>    <dbl>  <dbl>
## 1      2        0      0
## 2      2        0      0
## 3      1        0      0
## 4      1        0      0
## 5      2        0      0
## 6      2        0      0
## # … with 119,384 more rows
```

```
hotels %>%
  select(adults, children, bab
  arrange(-babies)
```

```
## # A tibble: 119,390 × 3
##   adults children babies
##    <dbl>    <dbl>  <dbl>
## 1      2        0     10
## 2      1        0      9
## 3      2        0      2
## 4      2        0      2
## 5      2        0      2
## 6      2        0      2
## # … with 119,384 more rows
```

filter

How could we remove the resort hotel? (`hotel` being `"Resort Hotel"`)

```
hotels %>%
  filter(___)
```

filter

How could we remove the resort hotel? (`hotel` being `"Resort Hotel"`)

```
hotels %>%
  filter(hotel != "Resort Hotel")
```

```
## # A tibble: 79,330 × 32
##   hotel      is_canceled lead_time arrival_date_ye… arrival_date_mo…
##   <chr>            <dbl>     <dbl>            <dbl> <chr>
## 1 City Hotel           0         6             2015 July
## 2 City Hotel           1        88             2015 July
## 3 City Hotel           1        65             2015 July
## 4 City Hotel           1        92             2015 July
## 5 City Hotel           1       100             2015 July
## 6 City Hotel           1        79             2015 July
## # … with 79,324 more rows, and 27 more variables:
## #   arrival_date_week_number <dbl>,
## #   arrival_date_day_of_month <dbl>,
## #   stays_in_weekend_nights <dbl>, stays_in_week_nights <dbl>,
## #   adults <dbl>, children <dbl>, babies <dbl>, meal <chr>,
## #   country <chr>, market_segment <chr>,
```

`filter` keeps rows matching conditions

Multiple conditions

Any bookings for only kids (no `adults`, but `children` or `babies`)?

Multiple conditions

Any bookings for only kids (no `adults`, but `children` or `babies`)?

```
hotels %>%
  filter(
    adults == 0,
    children >= 1 | babies >= 1     # | means or
    ) %>%
  select(adults, babies, children)
```

```
## # A tibble: 223 × 3
##   adults babies children
##    <dbl>  <dbl>    <dbl>
## 1      0      0        3
## 2      0      0        2
## 3      0      0        2
## 4      0      0        2
## 5      0      0        2
## 6      0      0        3
## # … with 217 more rows
```

Logical operators in R

operator	definition	operator	definition
`<`	less than	`x | y`	`x` OR `y`
`<=`	less than or equal to	`is.na(x)`	test if `x` is `NA`
`>`	greater than	`!is.na(x)`	test if `x` is not `NA`
`>=`	greater than or equal to	`x %in% y`	test if `x` is in `y`
`==`	exactly equal to	`!(x %in% y)`	test if `x` is not in `y`
`!=`	not equal to	`!x`	not `x`
`x & y`	`x` AND `y`		

Which market segments book most in advance?

market_segment	lead_time
Groups	186.973096
Offline TA/TO	135.004459
Online TA	82.998725
Direct	49.859115
Corporate	22.125590
Complementary	13.286676
Aviation	4.443038
Undefined	1.500000

How would we compute this?

Which market segments book most in advance?

market_segment	lead_time
Groups	186.973096
Offline TA/TO	135.004459
Online TA	82.998725
Direct	49.859115
Corporate	22.125590
Complementary	13.286676
Aviation	4.443038
Undefined	1.500000

```
hotels %>%
  group_by(market_segment) %>%
  summarize(
    lead_time = mean(lead_time))
```

```
## # A tibble: 8 × 2
##   market_segment lead_time
##   <chr>              <dbl>
## 1 Aviation            4.44
## 2 Complementary      13.3
## 3 Corporate          22.1
## 4 Direct             49.9
## 5 Groups            187.
## 6 Offline TA/TO     135.
## # … with 2 more rows
```

Which market segments book most in advance?

market_segment	lead_time
Groups	186.973096
Offline TA/TO	135.004459
Online TA	82.998725
Direct	49.859115
Corporate	22.125590
Complementary	13.286676
Aviation	4.443038
Undefined	1.500000

```
hotels %>%
  group_by(market_segment) %>%
  summarize(
    lead_time = mean(lead_time)) %>%
  arrange(desc(lead_time))
```

```
## # A tibble: 8 × 2
##   market_segment lead_time
##   <chr>              <dbl>
## 1 Groups             187.
## 2 Offline TA/TO      135.
## 3 Online TA           83.0
## 4 Direct              49.9
## 5 Corporate           22.1
## 6 Complementary       13.3
## # … with 2 more rows
```

How many total nights for each booking?

```
hotels %>%
  select(hotel, stays_in_week_nights, stays_in_weekend_nights)
```

```
## # A tibble: 119,390 × 3
##   hotel        stays_in_week_nights stays_in_weekend_nights
##   <chr>                       <dbl>                   <dbl>
## 1 Resort Hotel                    0                       0
## 2 Resort Hotel                    0                       0
## 3 Resort Hotel                    1                       0
## 4 Resort Hotel                    1                       0
## 5 Resort Hotel                    2                       0
## 6 Resort Hotel                    2                       0
## # … with 119,384 more rows
```

```
hotels %>%
  mutate(
    num_nights = ____)
```

How many total nights for each booking?

```
hotels %>%
  mutate(
    num_nights = stays_in_week_nights + stays_in_weekend_nights)
  select(hotel, num_nights)
```

```
## # A tibble: 119,390 × 2
##   hotel        num_nights
##   <chr>             <dbl>
## 1 Resort Hotel          0
## 2 Resort Hotel          0
## 3 Resort Hotel          1
## 4 Resort Hotel          1
## 5 Resort Hotel          2
## 6 Resort Hotel          2
## # … with 119,384 more rows
```

How long did each market segment stay?

How long did each market segment stay?

```
hotels %>%
  mutate(
    num_nights = stays_in_week_nights + stays_in_weekend_nights) %>%
  group_by(market_segment) %>%
  summarize(num_nights = mean(num_nights)) %>%
  arrange(desc(num_nights))
```

```
## # A tibble: 8 × 2
##   market_segment num_nights
##   <chr>                <dbl>
## 1 Offline TA/TO         3.90
## 2 Aviation              3.61
## 3 Online TA             3.57
## 4 Direct                3.21
## 5 Groups                2.99
## 6 Corporate             2.09
## # … with 2 more rows
```

How many in each market segment was a repeating guestguest?

```
hotels %>% distinct(is_repeated_guest)
```

```
## # A tibble: 2 × 1
##   is_repeated_guest
##               <dbl>
## 1                 0
## 2                 1
```

How many in each market segment was a repeating guestguest?

```
hotels %>% distinct(is_repeated_guest)
```

```
## # A tibble: 2 × 1
##   is_repeated_guest
##               <dbl>
## 1                 0
## 2                 1
```

```
hotels %>%
  group_by(market_segment) %>%
  summarize(frac_repeat = mean(is_repeated_guest)) %>%
  ggplot(aes(y = fct_reorder(market_segment, frac_repeat), x = f
    geom_col()
```