HW9

Andrew Shao

2024-11-06

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## 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 become errors
```

```
library(r02pro)
```

```
## Warning: package 'r02pro' was built under R version 4.3.3
```

```
d1 <- ahp %>%
    select(dt_sold, bsmt_area, bsmt_ht) %>%
    head(n = 5)
d2 <- tibble(bsmt_ht = c("Excellent", "Good", "Average", "Poor"), height = c("100+
↪   inches",
    "90-99 inches", "80-89 inches", "<70 inches"))
d1
```

```
## # A tibble: 5 x 3
##   dt_sold    bsmt_area bsmt_ht
##   <date>         <dbl> <chr>
## 1 2010-03-25       725 Average
## 2 2009-04-10       913 Good
## 3 2010-01-15      1057 Average
## 4 2010-04-19       384 Good
## 5 2010-03-22       676 Fair
```

```
#> # A tibble: 5 × 3
#>   dt_sold    bsmt_area bsmt_ht
#>   <date>         <dbl> <chr>
#> 1 2010-03-25       725 Average
#> 2 2009-04-10       913 Good
#> 3 2010-01-15      1057 Average
#> 4 2010-04-19       384 Good
#> 5 2010-03-22       676 Fair
d2
```

```
## # A tibble: 4 x 2
##   bsmt_ht   height
##   <chr>     <chr>
## 1 Excellent 100+ inches
## 2 Good      90-99 inches
## 3 Average   80-89 inches
## 4 Poor      <70 inches
```

```
#> # A tibble: 4 × 2
#>   bsmt_ht   height
#>   <chr>     <chr>
#> 1 Excellent 100+ inches
#> 2 Good      90-99 inches
#> 3 Average   80-89 inches
#> 4 Poor      <70 inches
```

7.7.7 Q1

```
inner_join(d1, d2, by = 'bsmt_ht')
```

```
## # A tibble: 4 x 4
##   dt_sold    bsmt_area bsmt_ht height
##   <date>         <dbl> <chr>   <chr>
## 1 2010-03-25       725 Average 80-89 inches
## 2 2009-04-10       913 Good    90-99 inches
## 3 2010-01-15      1057 Average 80-89 inches
## 4 2010-04-19       384 Good    90-99 inches
```

Only rows with matching keys in `d1` and `d2` are retained. All columns from `d1` and `d2` are retained.

7.7.7 Q2

```
left_join(d1, d2, by = 'bsmt_ht')
```

```
## # A tibble: 5 x 4
##   dt_sold    bsmt_area bsmt_ht height
##   <date>         <dbl> <chr>   <chr>
## 1 2010-03-25       725 Average 80-89 inches
## 2 2009-04-10       913 Good    90-99 inches
## 3 2010-01-15      1057 Average 80-89 inches
## 4 2010-04-19       384 Good    90-99 inches
## 5 2010-03-22       676 Fair    <NA>
```

All rows from d1 are retained and all rows from d2 with matching keys in d1 are retained. All columns from d1 and d2 are retained.

7.7.7 Q3

```
right_join(d1, d2, by = 'bsmt_ht')
```

```
## # A tibble: 6 x 4
##   dt_sold    bsmt_area bsmt_ht   height
##   <date>         <dbl> <chr>     <chr>
## 1 2010-03-25       725 Average   80-89 inches
## 2 2009-04-10       913 Good      90-99 inches
## 3 2010-01-15      1057 Average   80-89 inches
## 4 2010-04-19       384 Good      90-99 inches
## 5 NA                NA Excellent 100+ inches
## 6 NA                NA Poor      <70 inches
```

All rows from d2 are retained and all rows from d1 with matching keys in d2 are retained. All columns from d1 and d2 are retained.

7.7.7 Q4

```
full_join(d1, d2, by = 'bsmt_ht')
```

```
## # A tibble: 7 x 4
##   dt_sold    bsmt_area bsmt_ht   height
##   <date>         <dbl> <chr>     <chr>
## 1 2010-03-25       725 Average   80-89 inches
## 2 2009-04-10       913 Good      90-99 inches
## 3 2010-01-15      1057 Average   80-89 inches
## 4 2010-04-19       384 Good      90-99 inches
## 5 2010-03-22       676 Fair      <NA>
## 6 NA                NA Excellent 100+ inches
## 7 NA                NA Poor      <70 inches
```

All rows from d1 and d2 are retained, with NA filling in all the values when there is no matching key to join on. All columns from d1 and d2 are retained.

7.7.7 Q5

```
semi_join(d1, d2, by = 'bsmt_ht')
```

```
## # A tibble: 4 x 3
##   dt_sold    bsmt_area bsmt_ht
##   <date>         <dbl> <chr>
## 1 2010-03-25       725 Average
## 2 2009-04-10       913 Good
## 3 2010-01-15      1057 Average
## 4 2010-04-19       384 Good
```

The rows with `bsmt_ht` value of `'Good'` or `'Average'` are retained since they show up in both `d1` and `d2`. Only the columns from `d1` are retained.

7.7.7 Q6

```r
anti_join(d1, d2, by = 'bsmt_ht')
```

```
## # A tibble: 1 x 3
##   dt_sold    bsmt_area bsmt_ht
##   <date>         <dbl> <chr>
## 1 2010-03-22       676 Fair
```

Only the last row in `d1` is retained since the key value of `'Fair'` is the only one that shows up in `d1` but not `d2`, which is the criteria for inclusion. Only the columns from `d1` are retained.

7.7.7 Q7

```r
d2_new <- d2 %>% mutate(height_code = factor(d2$bsmt_ht, levels = c('Excellent', 'Good',
↪  'Average', 'Poor'), labels = c(1, 2, 3, 4)))
inner_join(d1, d2_new, by = 'bsmt_ht')
```

```
## # A tibble: 4 x 5
##   dt_sold    bsmt_area bsmt_ht height       height_code
##   <date>         <dbl> <chr>   <chr>        <fct>
## 1 2010-03-25       725 Average 80-89 inches 3
## 2 2009-04-10       913 Good    90-99 inches 2
## 3 2010-01-15      1057 Average 80-89 inches 3
## 4 2010-04-19       384 Good    90-99 inches 2
```

The new `height_code` column shows up in the results.

7.7.7 Q8

```r
d1_filter <- d1 %>% filter(bsmt_area > 600 & bsmt_area < 800)
inner_join(d1_filter, d2, by = 'bsmt_ht')
```

```
## # A tibble: 1 x 4
##   dt_sold    bsmt_area bsmt_ht height
##   <date>         <dbl> <chr>   <chr>
## 1 2010-03-25       725 Average 80-89 inches
```

1 row comes from `d2`.

7.7.7 Q9

```r
d1_na <- tibble(d1)
d1_na[1, 'bsmt_ht'] = NA
full_join(d1_na, d2, by = 'bsmt_ht')
```

```
## # A tibble: 7 x 4
##   dt_sold    bsmt_area bsmt_ht   height
##   <date>         <dbl> <chr>     <chr>
## 1 2010-03-25       725 <NA>      <NA>
## 2 2009-04-10       913 Good      90-99 inches
## 3 2010-01-15      1057 Average   80-89 inches
## 4 2010-04-19       384 Good      90-99 inches
## 5 2010-03-22       676 Fair      <NA>
## 6 NA                NA Excellent 100+ inches
## 7 NA                NA Poor      <70 inches
```

The missing value causes the first row to not be joined with any row from `d2` so its `height` value is `NA`.

7.7.7 Q10

```r
d2 <- rbind(d2, c('Very Good', '95-99 inches'))
anti_join(d1, d2, by = 'bsmt_ht')
```

```
## # A tibble: 1 x 3
##   dt_sold    bsmt_area bsmt_ht
##   <date>         <dbl> <chr>
## 1 2010-03-22       676 Fair
```

The rows in `d1` which don't have matching key values in `d2` are retained. Since the only `bsmt_ht` value which isn't in `d2` is `'Fair'` its row is the only one that is retained. Adding the last row to `d2` doesn't affect the result because its `bsmt_ht` value doesn't show up in `d1`.