Untitled

Jawaid Hakim

2022-09-28

Contents

The chart above describes arrival delays for two airlines across five destinations. Your task is to: (1) Create a .CSV file (or optionally, a MySQL database!) that includes all of the information above. You're encouraged to use a "wide" structure similar to how the information appears above, so that you can practice tidying and transformations as described below. (2) Read the information from your .CSV file into R, and use tidyr and dplyr as needed to tidy and transform your data. (3) Perform analysis to compare the arrival delays for the two airlines. (4) Your code should be in an R Markdown file, posted to rpubs.com, and should include narrative descriptions of your data cleanup work, analysis, and conclusions. Please include in your homework submission:

```
input_ds <- read_csv("airline_delays.csv")
```

```
## New names:
## Rows: 5 Columns: 7
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): ...1, ...2 dbl (5): Los Angeles, Phoenix, San Diego, San Francisco,
## Seattle
## 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.
## * ‘‘ -> ‘...1‘
## * ‘‘ -> ‘...2‘
```

```
str(input_ds)
```

```
## spec_tbl_df [5 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ ...1         : chr [1:5] "Alaska" NA NA "AM WEST" ...
##  $ ...2         : chr [1:5] "on time" "delayed" NA "on time" ...
##  $ Los Angeles  : num [1:5] 497 62 NA 694 117
##  $ Phoenix      : num [1:5] 221 12 NA 4840 415
##  $ San Diego    : num [1:5] 212 20 NA 383 65
##  $ San Francisco: num [1:5] 503 102 NA 320 129
##  $ Seattle      : num [1:5] 1841 305 NA 201 61
##  - attr(*, "spec")=
##   .. cols(
##   ..   ...1 = col_character(),
##   ..   ...2 = col_character(),
##   ..   ‘Los Angeles‘ = col_double(),
##   ..   Phoenix = col_double(),
##   ..   ‘San Diego‘ = col_double(),
##   ..   ‘San Francisco‘ = col_double(),
##   ..   Seattle = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
input_ds <- input_ds %>% rename(airline = 1, arrival_status = 2)
```

```
str(input_ds)
```

```
## spec_tbl_df [5 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ airline       : chr [1:5] "Alaska" NA NA "AM WEST" ...
##  $ arrival_status: chr [1:5] "on time" "delayed" NA "on time" ...
##  $ Los Angeles   : num [1:5] 497 62 NA 694 117
##  $ Phoenix       : num [1:5] 221 12 NA 4840 415
##  $ San Diego     : num [1:5] 212 20 NA 383 65
##  $ San Francisco : num [1:5] 503 102 NA 320 129
##  $ Seattle       : num [1:5] 1841 305 NA 201 61
##  - attr(*, "spec")=
##   .. cols(
##   ..   ...1 = col_character(),
##   ..   ...2 = col_character(),
##   ..   ‘Los Angeles‘ = col_double(),
##   ..   Phoenix = col_double(),
##   ..   ‘San Diego‘ = col_double(),
##   ..   ‘San Francisco‘ = col_double(),
##   ..   Seattle = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
head(input_ds)
```

```
## # A tibble: 5 x 7
##   airline arrival_status ‘Los Angeles‘ Phoenix ‘San Diego‘ San Francis~1 Seattle
##   <chr>   <chr>                  <dbl>   <dbl>       <dbl>         <dbl>   <dbl>
## 1 Alaska  on time                  497     221         212           503    1841
## 2 <NA>    delayed                   62      12          20           102     305
## 3 <NA>    <NA>                      NA      NA          NA            NA      NA
## 4 AM WEST on time                  694    4840         383           320     201
## 5 <NA>    delayed                  117     415          65           129      61
## # ... with abbreviated variable name 1: ‘San Francisco‘
```

```
input_ds <- input_ds %>%
            filter(! is.na(arrival_status)) %>%
            fill(airline)
head(input_ds)
```

```
## # A tibble: 4 x 7
##   airline arrival_status ‘Los Angeles‘ Phoenix ‘San Diego‘ San Francis~1 Seattle
##   <chr>   <chr>                  <dbl>   <dbl>       <dbl>         <dbl>   <dbl>
## 1 Alaska  on time                  497     221         212           503    1841
## 2 Alaska  delayed                   62      12          20            102     305
## 3 AM WEST on time                  694    4840         383            320     201
## 4 AM WEST delayed                  117     415          65            129      61
## # ... with abbreviated variable name 1: `San Francisco`
```

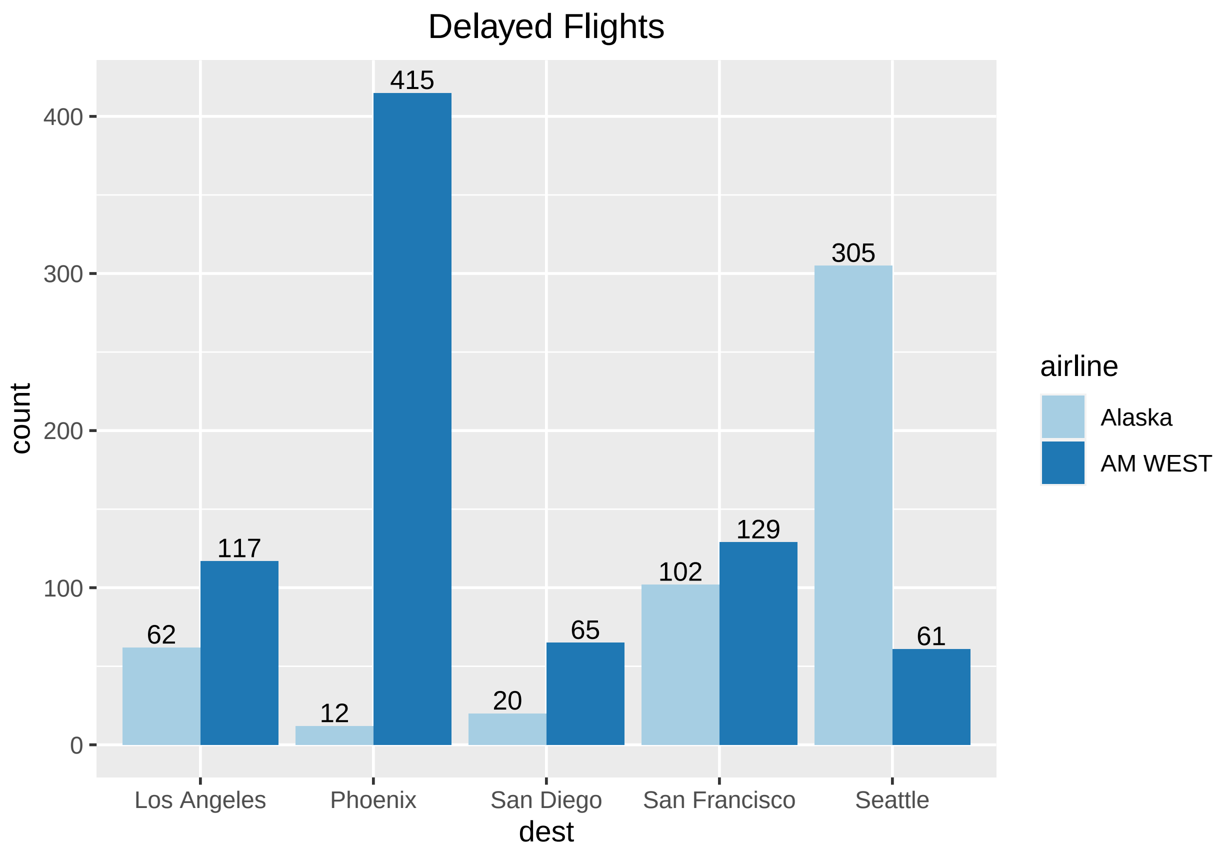
```
input_ds <- input_ds %>%
                pivot_longer(!c("airline", "arrival_status"),
                            names_to = "dest",
                            values_to = "count")
head(input_ds)
```

```
## # A tibble: 6 x 4
##   airline arrival_status dest          count
##   <chr>   <chr>          <chr>         <dbl>
## 1 Alaska  on time        Los Angeles     497
## 2 Alaska  on time        Phoenix         221
## 3 Alaska  on time        San Diego       212
## 4 Alaska  on time        San Francisco   503
## 5 Alaska  on time        Seattle        1841
## 6 Alaska  delayed        Los Angeles      62
```

```
delayed_flights <- input_ds %>%
                    filter(input_ds$arrival_status == "delayed")
delayed_flights
```

```
## # A tibble: 10 x 4
##    airline arrival_status dest          count
##    <chr>   <chr>          <chr>         <dbl>
##  1 Alaska  delayed        Los Angeles      62
##  2 Alaska  delayed        Phoenix          12
##  3 Alaska  delayed        San Diego        20
##  4 Alaska  delayed        San Francisco   102
##  5 Alaska  delayed        Seattle         305
##  6 AM WEST delayed        Los Angeles     117
##  7 AM WEST delayed        Phoenix         415
##  8 AM WEST delayed        San Diego        65
##  9 AM WEST delayed        San Francisco   129
## 10 AM WEST delayed        Seattle          61
```

```
    ggp <- ggplot(data=delayed_flights, aes(x=dest, y=count, fill=airline))
    ggp <- ggp +  ggtitle('Delayed Flights') + theme(plot.title = element_text(hjust = 0.5))
    ggp <- ggp + geom_text(aes(label=count), vjust=-0.2,
                            position = position_dodge(0.9), size=3.5) +
                            scale_fill_brewer(palette="Paired") +
            geom_bar(stat="identity", position=position_dodge())
ggp
```


Delayed Flights
count
400
300
200
100
0
415
305
117
129
102
62
65
61
12
20
Los Angeles
Phoenix
San Diego
San Francisco
Seattle
dest
airline
Alaska
AM WEST