Data_606_Assignment_4_Working_with_Tidy_Data.Rmd

Enid Roman

2022-10-04

```
# Upload the libraries needed.

library(tidyr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v dplyr   1.0.9
## v tibble  3.1.8     v stringr 1.4.1
## v readr   2.1.2     v forcats 0.5.2
## v purrr   0.3.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(ggplot2)
```

```
# Import the data from github.
# Link is provided to the csv file below:
# https://github.com/enidroman/data_607_data_acquisition_and_management

urlfile <- "https://raw.githubusercontent.com/enidroman/data_607_data_acquisition_and_management/main/T

table <- read.csv(urlfile)
table
```

```
##         X     X.1 Los.Angeles Phoenix San.Diego San.Francisco Seattle
## 1  ALASKA on time         497     221       212           503   1,841
## 2         delayed          62      12        20           102     305
## 3                          NA              NA            NA
## 4 AM WEST on time         694   4,840       383           320     201
## 5         delayed         117     415        65           129      61
```

DATA CLEANING AND TRANSFORMATION

In observing the table I see that:

1. The first 2 columns needs to be renamed, X = Airline, X.1 = Status
2. There is a blank row that separates the airlines that needs to be removed.
3. The airline names needs be brought down to be aligned with the delayed.

4. Both columns, Phoenix and Seattle, are characters instead of integers. Commas from the numbers 4,480 and 1,840 have to be removed in order to convert the columns Phoenix and Seattle from character to integer.
5. Each variable in the dataset should have its own column. The cities are listed as separate columns when they should be combined into 1 variable. The X.1 = status column contains values that should be split into 2 separate variables.
6. The period (.) between Los Angeles, San Diego, and San Francisco needs to be replaced by a space.

```
table2 <- table %>%
  rename(AIRLINE = X, STATUS = X.1)    # Renamed column X = AIRLINE and X.1 = STATUS

table2 <- drop_na(table2)              # Removed blank row that seperates the airlines.

table2[table2==""]<-NA                 # Bring down the Airlines name  to be aligned with the Status Col
table2 <- fill(table2, AIRLINE)

table2
```

```
##   AIRLINE  STATUS Los.Angeles Phoenix San.Diego San.Francisco Seattle
## 1  ALASKA on time         497     221       212           503   1,841
## 2  ALASKA delayed          62      12        20           102     305
## 3 AM WEST on time         694   4,840       383           320     201
## 4 AM WEST delayed         117     415        65           129      61
```

```
table2$Phoenix <- as.integer(gsub(",","",table2$Phoenix))   # Removed comma in 4,480 in Phoenix to conv
sapply(table2, class)
```

```
##       AIRLINE        STATUS   Los.Angeles       Phoenix     San.Diego
##   "character"   "character"     "integer"     "integer"     "integer"
## San.Francisco       Seattle
##     "integer"   "character"
```

```
table2$Seattle <- as.integer(gsub(",","",table2$Seattle))   # Removed comma in 1,841 in Seattle to conv
sapply(table2, class)
```

```
##       AIRLINE        STATUS   Los.Angeles       Phoenix     San.Diego
##   "character"   "character"     "integer"     "integer"     "integer"
## San.Francisco       Seattle
##     "integer"     "integer"
```

```
# Combined all city in City Column while aligning the cities with the airline names. Created a Delayed

table2 <- table2 %>%
  gather(CITY,NUM_FLIGHTS, -AIRLINE, -STATUS) %>%
  spread(STATUS, NUM_FLIGHTS)

colnames(table2) <- c('AIRLINE', 'CITY', 'DELAYED', 'ON_TIME')

table2$CITY <- str_replace_all(table2$CITY, "\\.", " ") # Replaced "," between the cities with a space.

table2
```

```
##    AIRLINE          CITY DELAYED ON_TIME
## 1   ALASKA   Los Angeles      62     497
## 2   ALASKA       Phoenix      12     221
## 3   ALASKA     San Diego      20     212
## 4   ALASKA San Francisco     102     503
## 5   ALASKA       Seattle     305    1841
## 6  AM WEST   Los Angeles     117     694
## 7  AM WEST       Phoenix     415    4840
## 8  AM WEST     San Diego      65     383
## 9  AM WEST San Francisco     129     320
## 10 AM WEST       Seattle      61     201
```

ANALYSIS

```
table3 <- table2                              # Created a dataframe with just Airline, City, and Delayed
select(table2, AIRLINE, CITY, DELAYED)
```

Analysis to compare the arrival delays for the two airlines.

```
##    AIRLINE          CITY DELAYED
## 1   ALASKA   Los Angeles      62
## 2   ALASKA       Phoenix      12
## 3   ALASKA     San Diego      20
## 4   ALASKA San Francisco     102
## 5   ALASKA       Seattle     305
## 6  AM WEST   Los Angeles     117
## 7  AM WEST       Phoenix     415
## 8  AM WEST     San Diego      65
## 9  AM WEST San Francisco     129
## 10 AM WEST       Seattle      61
```

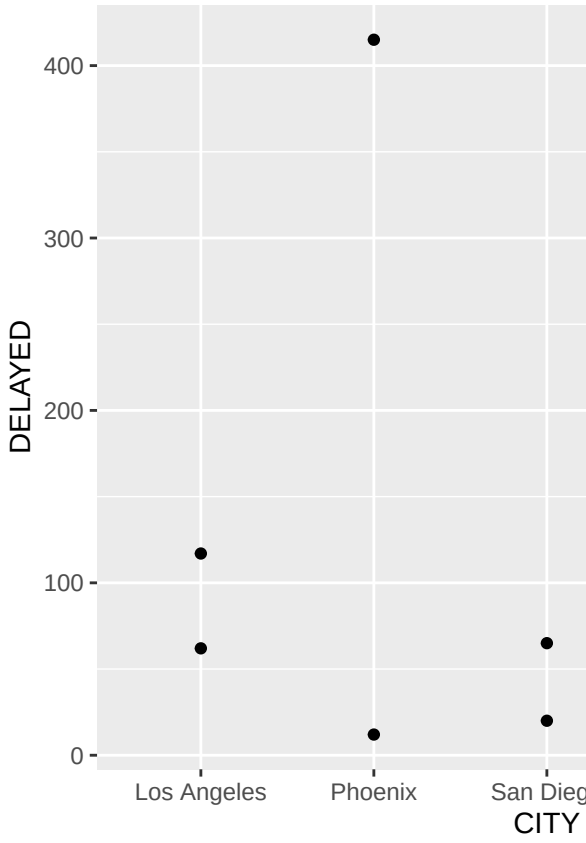
```
flights_delayed <- table3
```

Note: I tried to do a geom point graph of the delays per city for each airline but I was unsuccessful in adding color to the dots to distinguish each airline.

```
#ggplot(data = flights_delayed) +
 # geom_bar(mapping = aes(x = CITY, y = DELAYED, fill = AIRLINE), position = "dodge")

ggplot(data = flights_delayed, mapping = aes(x = CITY, y = DELAYED, fill = AIRLINE), position = "dodge"
  geom_point()
```

This graph is representing the above data delayed for each airline by cities. The longest delay



was with AM West with Phoenix and with Alsaska was Seattle.

Here you can see that AM WEST had 286 more delays then Alaska. That is an average of 57.2 more then Alaska. AM WEST had a median of 117 delays vs 62 delays of Alaska and a minimum of 61 delays vs 12 delays of Alaska, and max of 415 delays vs 305 delays of Alaska. More investigation has to be done to find the reason behind the delays in AM West.

```
flights_delayed %>%                          # Summary of delays for each airline.
  group_by(AIRLINE) %>%
  summarise(TOTAL_DELAYS = sum(DELAYED),     # Total sum of delays for each airline.
            AVG_NUM_DELAYS = mean(DELAYED),  # Average of delays for each airline.
            MEDIAN_DELAYS = median(DELAYED), # The median of delays for each airline.
            MIN_DELAYS = min(DELAYED),       # The minimum of delays for each airline.
            MAX_DELAYS = max(DELAYED))       # The maximum of delays for each airline.
```

```
## # A tibble: 2 x 6
##   AIRLINE TOTAL_DELAYS AVG_NUM_DELAYS MEDIAN_DELAYS MIN_DELAYS MAX_DELAYS
##   <chr>          <int>          <dbl>         <int>      <int>      <int>
## 1 ALASKA           501           100.            62         12        305
## 2 AM WEST          787           157.           117         61        415
```