Wk 5: Tidying and Transforming Data

Dirk Hartog

2023-10-01

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.4
## v ggplot2   3.4.3     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0

## -- 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
```

Assignment – Tidying and Transforming Data

1. Created a .CSV file in Google sheets and uploaded it into a Github repository to be laoded into Rstudio

2a Read the information from the .CSV file into R

```
url <- "https://raw.githubusercontent.com/D-hartog/DATA607/main/airline_status.csv"
airline_info <- read_csv(url)
```

```
## 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`
```

```
glimpse(airline_info)
```

```
## Rows: 5
## Columns: 7
## $ ...1            <chr> "ALASKA", NA, NA, "AM WEST", NA
## $ ...2            <chr> "on_time", "delayed", NA, "on_time", "delayed"
## $ `Los Angeles`   <dbl> 497, 62, NA, 694, 117
## $ Phoenix         <dbl> 221, 12, NA, 4840, 415
## $ `San Diego`     <dbl> 212, 20, NA, 383, 65
## $ `San Francisco` <dbl> 503, 102, NA, 320, 129
## $ Seattle         <dbl> 1841, 305, NA, 201, 61
```

2b. Used tidyr and dplyr as needed to tidy the data

```
# Renamed columns using rename() function

airline_info <- airline_info %>%
  rename("airline" = "...1",
         "status" = "...2",
         "Los_Angeles" = "Los Angeles",
         "San_Diego" = "San Diego",
         "San_Francisco" = "San Francisco")

# Dropped any rows with all NA values
airline_info <- airline_info %>%
  filter(rowSums(is.na(airline_info)) != ncol(airline_info))

# Fill in the NA values in the "airline" column
airline_info[2,"airline"] <- airline_info[1,"airline"]
airline_info[4,"airline"] <- airline_info[3,"airline"]
```

2c. Transformed table

```
# pivot the table into a longer format by moving the city columns to value and creating a new count col

airline_info <- airline_info %>%
  pivot_longer(
    cols = Los_Angeles:Seattle,
```

```
    names_to = "dest",
    values_to = "count"
  )

# Then pivot the status column into two new columns using the respective count values as values
airline_info <- airline_info %>%
  pivot_wider(
    names_from = status,
    values_from = count
  )

airline_info
```

```
## # A tibble: 10 x 4
##    airline dest          on_time delayed
##    <chr>   <chr>           <dbl>   <dbl>
##  1 ALASKA  Los_Angeles       497      62
##  2 ALASKA  Phoenix           221      12
##  3 ALASKA  San_Diego         212      20
##  4 ALASKA  San_Francisco     503     102
##  5 ALASKA  Seattle          1841     305
##  6 AM WEST Los_Angeles       694     117
##  7 AM WEST Phoenix          4840     415
##  8 AM WEST San_Diego         383      65
##  9 AM WEST San_Francisco     320     129
## 10 AM WEST Seattle           201      61
```

3. Perform analysis to compare the arrival delays for the two airlines.

Descriptive statistics of delays

```
airline_info %>%
  group_by(airline) %>%
  summarise(Mean = mean(delayed),
            Median = median(delayed),
            IQR = IQR(delayed),
            Maximum = max(delayed),
            Minimum = min(delayed))
```

```
## # A tibble: 2 x 6
##   airline  Mean Median   IQR Maximum Minimum
##   <chr>   <dbl>  <dbl> <dbl>   <dbl>   <dbl>
## 1 ALASKA   100.     62    82     305      12
## 2 AM WEST  157.    117    64     415      61
```

```
airline_info %>%
  group_by(dest) %>%
  summarise(Average = mean(delayed),
            Maximum = max(delayed),
            Minimum = min(delayed))
```

```
## # A tibble: 5 x 4
##   dest          Average Maximum Minimum
##   <chr>           <dbl>   <dbl>   <dbl>
## 1 Los_Angeles      89.5     117      62
## 2 Phoenix         214.      415      12
## 3 San_Diego        42.5      65      20
## 4 San_Francisco   116.      129     102
## 5 Seattle         183       305      61
```

Compare the average proportion of delayed flights between the two airlines

```
# Find the proportion of delays from each airline and the destination
airline_info <- airline_info %>%
  mutate(pct_delayed = (delayed/(delayed + on_time)))
```

It might be interesting to track this overtime to see any trends in the delays overtime

Summarizing the average number of flights and average percent of delays by airline

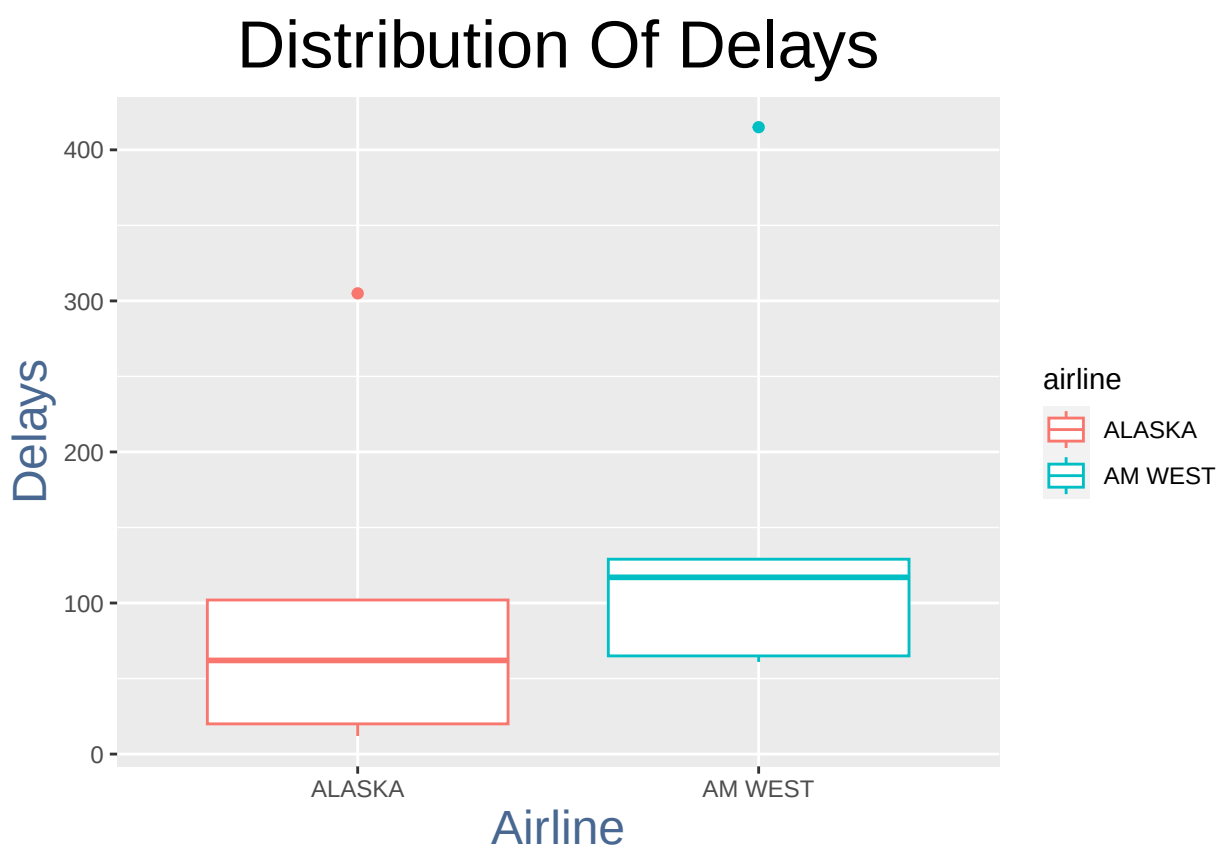
```
airline_info %>%
  group_by(airline) %>%
  summarize(Avg_delyed_flights = mean(delayed),
            Avg_percent_delayed = mean(pct_delayed))
```

```
## # A tibble: 2 x 3
##   airline Avg_delyed_flights Avg_percent_delayed
##   <chr>                <dbl>               <dbl>
## 1 ALASKA                100.               0.112
## 2 AM WEST               157.               0.178
```

4. Visualizations

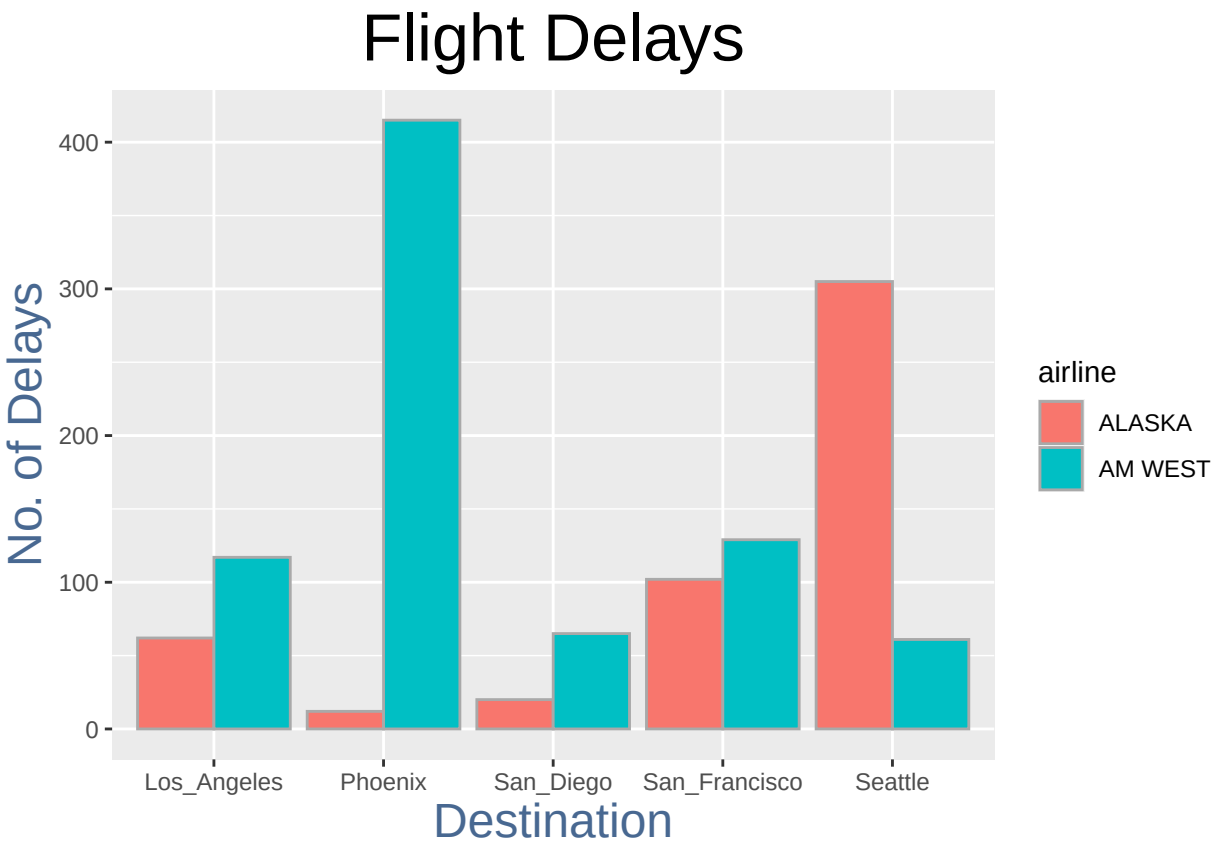
Visualization of the distribution of the data via box plot of number of flights on time and the delayed flights

```
ggplot(data = airline_info,
       mapping = aes(y = delayed, x = airline, color = airline)) +
  geom_boxplot() +
  ylab("Delays") +
  xlab("Airline") +
  ggtitle("Distribution Of Delays") +
  theme(axis.title.y = element_text(color="#486891", size=18),
        axis.title.x = element_text(color="#486891", size=18),
        plot.title = element_text(color="black",
                                  size=25,
                                  hjust = 0.5))
```



Bar plot of the counts based on airline and destination

```
ggplot(data = airline_info, aes(dest,delayed)) +
  geom_col(aes(color = airline, fill = airline),
           position = "dodge", color = "darkgrey") +
  ylab("No. of Delays") +
  xlab("Destination") +
  ggtitle("Flight Delays") +
  theme(axis.title.y = element_text(color="#486891",
                                    size=18),
        axis.title.x = element_text(color="#486891",
                                    size=18),
        plot.title = element_text(color="black",
                                  size=25,
                                  hjust = 0.5))
```


Flight Delays
No. of Delays
400
300
200
100
0
Los_Angeles
Phoenix
San_Diego
San_Francisco
Seattle
Destination
airline
ALASKA
AM WEST