Flights Tidiying

MGinorio

Assignment – Tidying and Transforming Data

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

Source: *Numbersense*, Kaiser Fung, McGraw Hill, 2013

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:

- The URL to the .Rmd file in your GitHub repository. and
- The URL for your rpubs.com web page.

Process

Overview

Pending

Packages

```
#Packages used

library(tidytext)
library(tidyverse)
library(ggplot2)
library(dplyr)
library(plotly)
library(stringr)
```

Dataset

Names Notice that because we used read_csv(), the data frame we received now prints nicely without having to use the head() function and does not clutter your screen.

```
flights <-read.csv("https://raw.githubusercontent.com/mgino11/Flights_Mani/main/HW_5_flights.csv",string

glimpse(flights)
```

```
## Rows: 4
## Columns: 7
## $ X             <chr> "ALASKA", "ALASKA", "AM WEST", "AM WEST"
## $ X.1           <chr> "on time", "delayed", "on time", "delayed"
## $ Los.Angeles   <int> 497, 62, 694, 117
## $ Phoenix       <int> 221, 12, 4840, 415
## $ San.Diego     <int> 212, 20, 383, 65
## $ San.Francisco <int> 503, 102, 320, 129
## $ Seattle       <int> 1841, 305, 201, 61
```

We need to get rid of spaces in "on time" so we can later manipulate data

```
flights[,2] <- sapply(flights[,2], str_replace, " ", "_")
flights
```

```
##         X     X.1 Los.Angeles Phoenix San.Diego San.Francisco Seattle
## 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
```

Manipulate

Clean data. Add the header Airline and Arrival Status to column 1 and 2. Pivot Longer for Arrival City

```
flights_pivot_1 <-flights %>%
  rename(airline = X, arrival.status = X.1) %>%
   pivot_longer(flights,
    cols = c(Los.Angeles, Phoenix, San.Diego, San.Francisco, Seattle),
    names_to = "Arrival_City",
    values_to = "Flight"
    )
```

```
## Warning in gsub(paste0("^", names_prefix), "", names(cols)): argument 'pattern'
## has length > 1 and only the first element will be used
```

```
flights_pivot_1
```

```
## # A tibble: 20 x 4
##    airline arrival.status Arrival_City  Flight
##    <chr>   <chr>          <chr>          <int>
##  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
##  7 ALASKA  delayed        Phoenix           12
##  8 ALASKA  delayed        San.Diego         20
##  9 ALASKA  delayed        San.Francisco    102
## 10 ALASKA  delayed        Seattle          305
## 11 AM WEST on_time        Los.Angeles      694
## 12 AM WEST on_time        Phoenix         4840
## 13 AM WEST on_time        San.Diego        383
## 14 AM WEST on_time        San.Francisco    320
## 15 AM WEST on_time        Seattle          201
## 16 AM WEST delayed        Los.Angeles      117
## 17 AM WEST delayed        Phoenix          415
## 18 AM WEST delayed        San.Diego         65
## 19 AM WEST delayed        San.Francisco    129
## 20 AM WEST delayed        Seattle           61
```

Pivot Wider I want to know the airline the Arrival City and what flight is on time or delayed

```
flights_pivot_1 <- flights_pivot_1 %>%
  pivot_wider(names_from = arrival.status,
              values_from = Flight)

flights_pivot_1
```

```
## # A tibble: 10 x 4
##    airline Arrival_City  on_time delayed
##    <chr>   <chr>           <int>   <int>
##  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
```

```
flights_pivot_1 <- flights_pivot_1 %>%
  mutate(total_flights = on_time + delayed,
         delayed_avg =   round((delayed/total_flights)*100,2)) %>%
  arrange(desc(delayed_avg))

flights_pivot_1
```

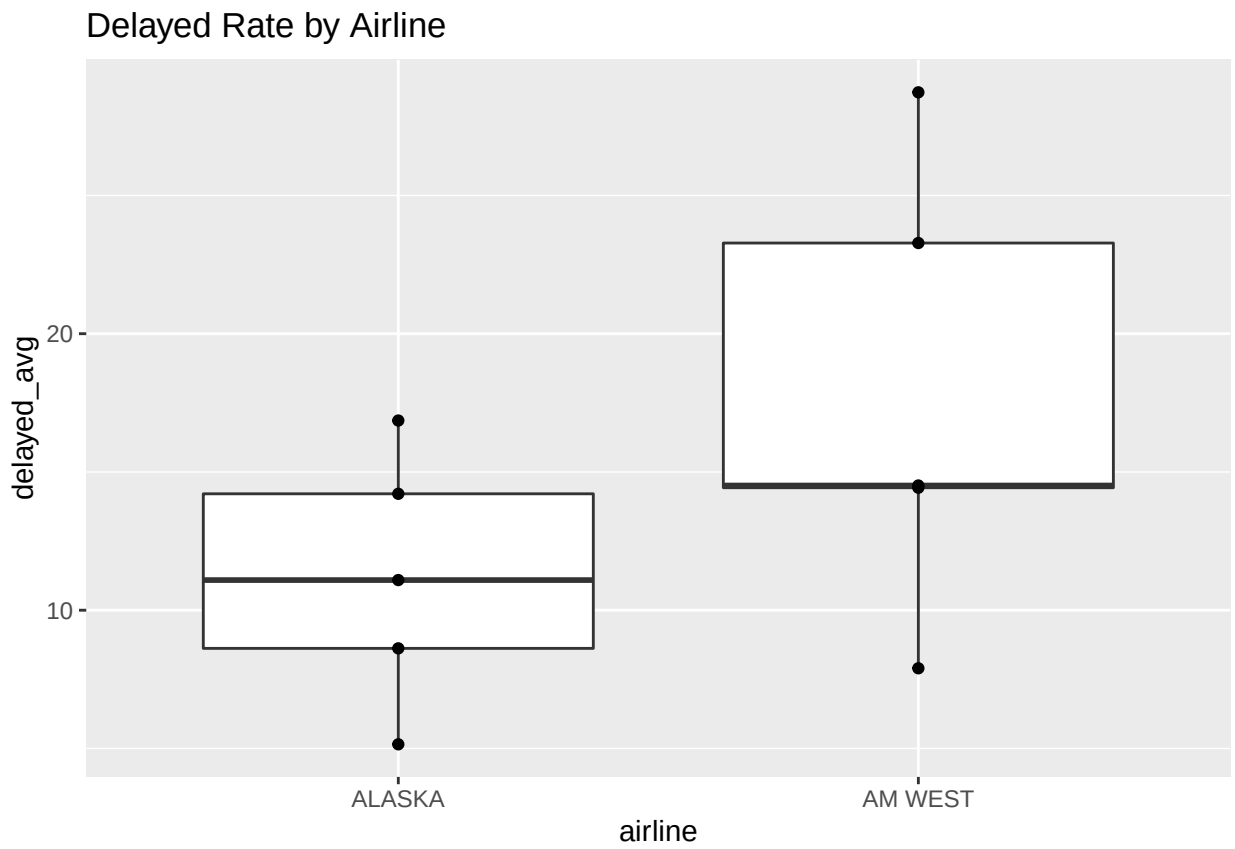
Transform

```
## # A tibble: 10 x 6
##    airline Arrival_City  on_time delayed total_flights delayed_avg
##    <chr>   <chr>           <int>   <int>         <int>       <dbl>
##  1 AM WEST San.Francisco     320     129           449       28.7
##  2 AM WEST Seattle           201      61           262       23.3
##  3 ALASKA  San.Francisco     503     102           605       16.9
##  4 AM WEST San.Diego         383      65           448       14.5
##  5 AM WEST Los.Angeles       694     117           811       14.4
##  6 ALASKA  Seattle          1841     305          2146       14.2
##  7 ALASKA  Los.Angeles       497      62           559       11.1
##  8 ALASKA  San.Diego         212      20           232        8.62
##  9 AM WEST Phoenix          4840     415          5255        7.9
## 10 ALASKA  Phoenix           221      12           233        5.15
```

Graphs

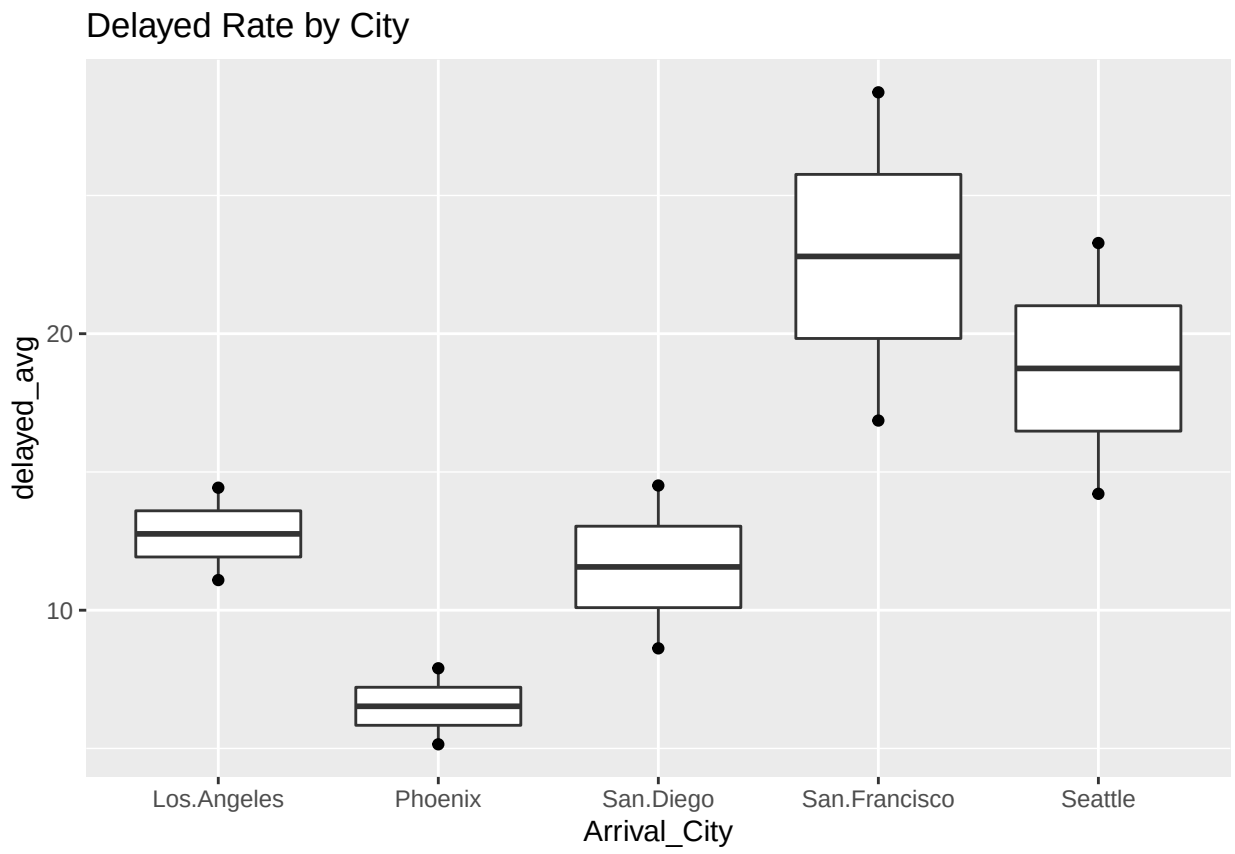
Delayed Avg by Airline I want to interpret graphically the frequency of delayed flights by airline

```
ggplot(flights_pivot_1, aes(x = airline, y = delayed_avg,)) +
  geom_boxplot() + geom_point() + ggtitle("Delayed Rate by Airline")
```



Delayed Avg by Airline I want to interpret graphically the frequency of delayed flights by by City

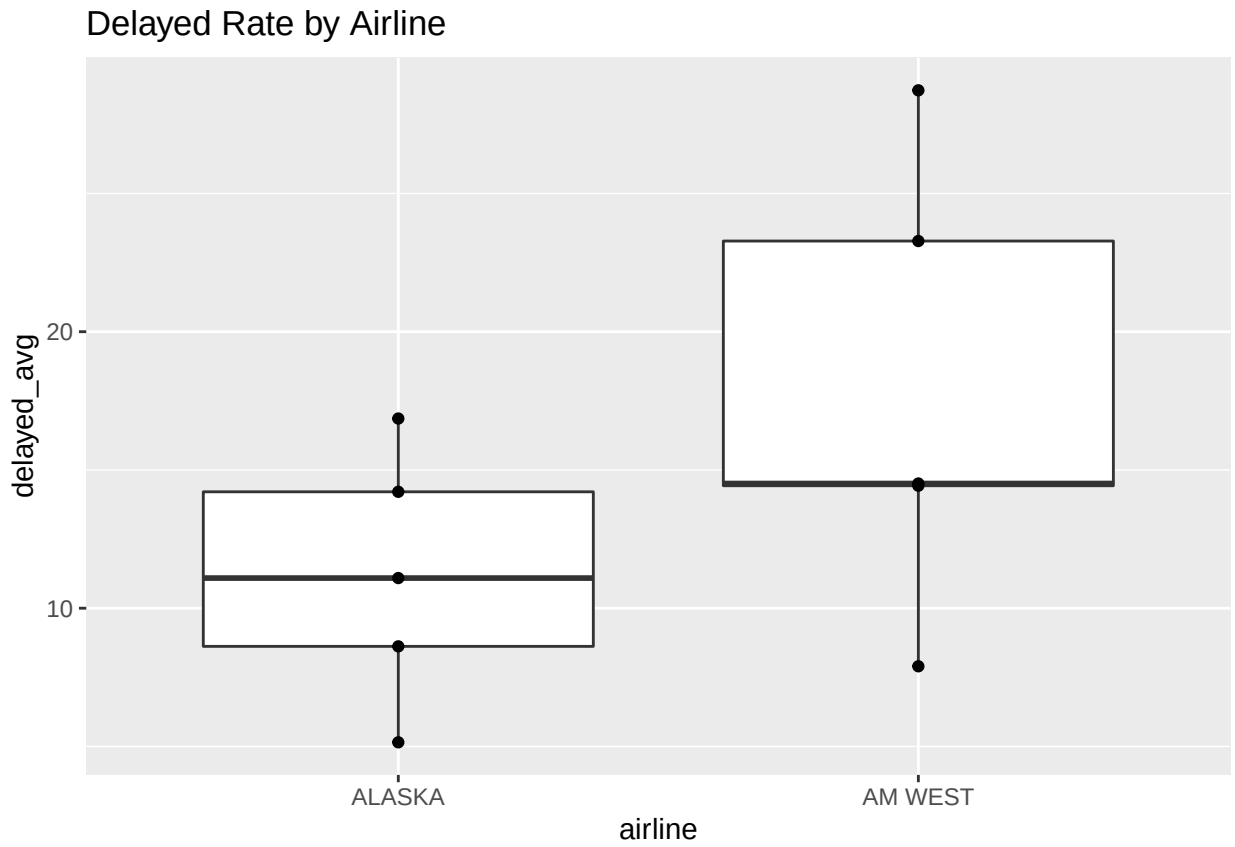
```
ggplot(flights_pivot_1, aes(x = Arrival_City, y = delayed_avg)) +
  geom_boxplot() + geom_point() + ggtitle("Delayed Rate by City")
```



Conclusion

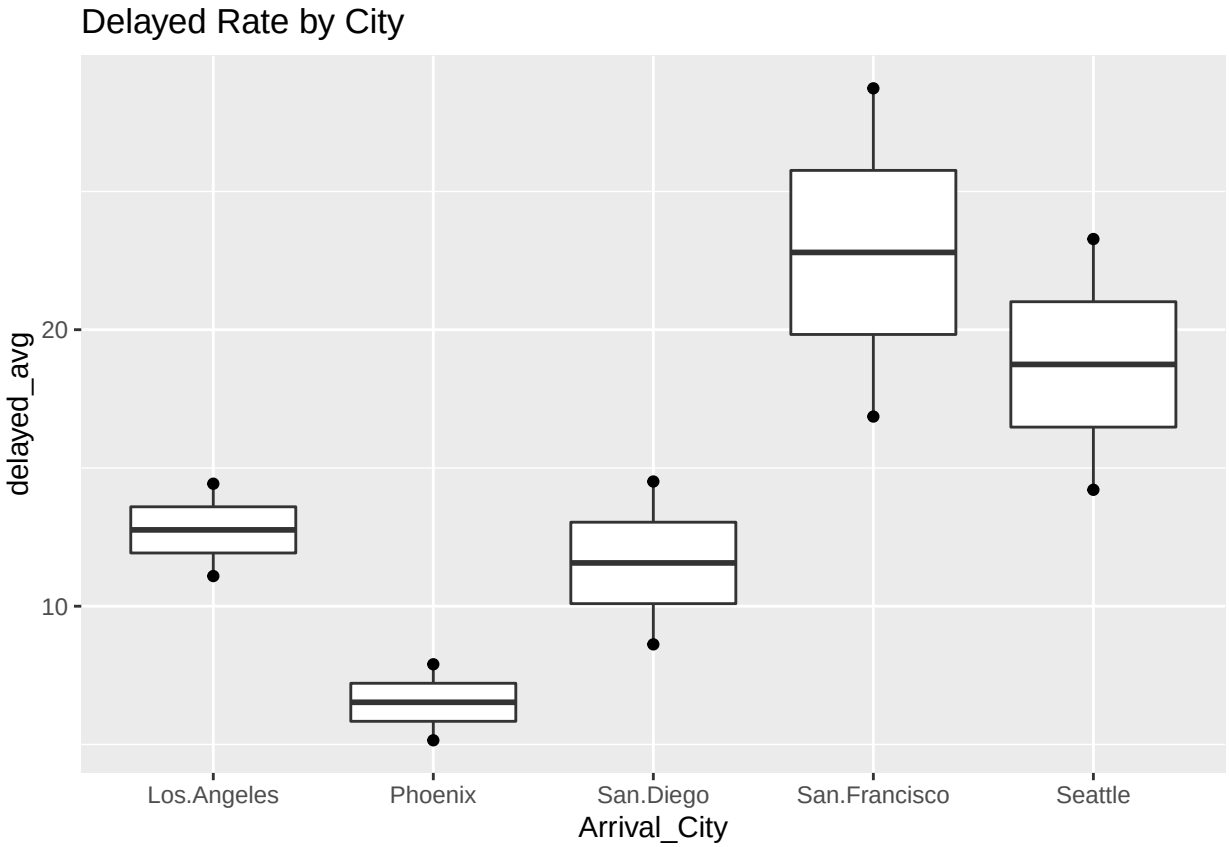
Delays By Airline Thanks to the data Wrangling we were able to determine that AM West has more delays compared to Alaska Airlines

```
ggplot(flights_pivot_1, aes(x = airline, y = delayed_avg,)) +
  geom_boxplot() + geom_point() + ggtitle("Delayed Rate by Airline")
```



Delays By city Thanks to the data wrangling I was able to determine that SFO and SEATTLE are the two cities with the most delays on average for the two airlines considered in this data Set.

```
ggplot(flights_pivot_1, aes(x = Arrival_City, y = delayed_avg)) +
  geom_boxplot() + geom_point() + ggtitle("Delayed Rate by City")
```


Delayed Rate by City
delayed_avg
20
10
Los.Angeles
Phoenix
San.Diego
San.Francisco
Seattle
Arrival_City