R Notebook

Neil Shah: DATA 607: HW 5

Introduction

The purpose of this assignment is to explore the Tidyverse universe/methodology on cleaning data for better analysis. Most data scientists spend the bulk of their time manipulating data for processing, and hence this is a very important skill.

Loading Relevant Packages

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

Loading Data

I manually inputed the data into a .csv file [flights.csv] and uploaded it both to my GitHub and to a local disk.

```
> flights <- read.csv("https://raw.githubusercontent.com/shahneilp/DATA607/master/HW5/flights.cs
v")
> head(flights)
      ï..       X Los.Angeles Phoenix San.Diego San.Francisco Seattle
1  Alaska on time         497     221       212           503    1841
2         delayed          62      12        20           102     305
3                          NA      NA        NA            NA      NA
4 AM West on time         694    4840       383           320     201
5         delayed         117     415        65           129      61
```

Cleaning out NAs

We have a few NA values due to the space in row 3; If this was a larger data-set, a more thorough analysis would involve summing NA values by columns, comparing it to length vectors and then getting an idea of what % is NA– this way you can see how much data you are manipulating.

Since this is a small table–and I visually know the NA's are just table artifacts, I can drop them.

```
> flights <- flights %>% drop_na()
> flights
      ï..       X Los.Angeles Phoenix San.Diego San.Francisco Seattle
1  Alaska on time         497     221       212           503    1841
2         delayed          62      12        20           102     305
4 AM West on time         694    4840       383           320     201
5         delayed         117     415        65           129      61
```

And now to reset the index

```
rownames(flights) <- 1:4
> flights
      ï..       X Los.Angeles Phoenix San.Diego San.Francisco Seattle
1  Alaska on time         497     221       212           503    1841
2         delayed          62      12        20           102     305
3 AM West on time         694    4840       383           320     201
4         delayed         117     415        65           129      61
```

Renaming Columns

The first two columns don't have names that align with their data–this is common when you load data without ordered headers. Also let's clean up the names of the cities [the periods ins tead of white space]

```
> names(flights) <- c('Airlines', 'Status','Los Angeles','Phoenix','San Diego','San Francisc
o','Seattle')
> flights
  Airlines  Status Los Angeles Phoenix San Diego San Francisco Seattle
1   Alaska on time         497     221       212           503    1841
2          delayed          62      12        20           102     305
3  AM West on time         694    4840       383           320     201
4          delayed         117     415        65           129      61
```

Filling out rows

Finally, we should make sure that our flight dataframe has consistent labels under the "delayed" row for Airlines. There are multiple ways to do this–if this was a larger data set I can iterate through i-1 rows (given it's just every other row). Once again this is a small dataframe, so I can do it manually to prevent errors.

```
> flights$Airlines[2] = 'Alaska'
> flights$Airlines[4] = 'AM West'
> flights
  Airlines  Status 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
```

Tidy Dataframe

The next step is to convert this wide table to a tidy dataframe–I will use gather to organize the data.

```
> flights <- gather(flights,"City","Count",3:7)
> flights
   Airlines  Status          City Count
1    Alaska on time   Los Angeles   497
2    Alaska delayed   Los Angeles    62
3   AM West on time   Los Angeles   694
4   AM West delayed   Los Angeles   117
5    Alaska on time       Phoenix   221
6    Alaska delayed       Phoenix    12
7   AM West on time       Phoenix  4840
8   AM West delayed       Phoenix   415
9    Alaska on time     San Diego   212
10   Alaska delayed     San Diego    20
11  AM West on time     San Diego   383
12  AM West delayed     San Diego    65
13   Alaska on time San Francisco   503
14   Alaska delayed San Francisco   102
15  AM West on time San Francisco   320
16  AM West delayed San Francisco   129
17   Alaska on time       Seattle  1841
18   Alaska delayed       Seattle   305
19  AM West on time       Seattle   201
20  AM West delayed       Seattle    61
```

Much better–but I think it'd be interesting to see the spread between delayed/on-time, that way we can make performance metrics.

```
> flights <- spread(flights,Status,Count)
> flights
   Airlines          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
```

Awesome–now time for analysis.

Performance Metrics

We want to compare delayed versus on time for cities and by airlines; I need to come up with a performance metric that best captures this–I think a simple % on time and % delayed would work!

I'll make a new column for "Delayed Ratio" "On Time Ratio"–note total flights would be the SUM of these columns.

```
> flights$DelayedRatio <- (flights$delayed/(flights$delayed+flights$`on time`))
> flights$OnTimeRatio <- (flights$`on time`/(flights$delayed+flights$`on time`))
> flights
   Airlines          City delayed on time DelayedRatio OnTimeRatio
1    Alaska   Los Angeles      62     497   0.11091234   0.8890877
2    Alaska       Phoenix      12     221   0.05150215   0.9484979
3    Alaska     San Diego      20     212   0.08620690   0.9137931
4    Alaska San Francisco     102     503   0.16859504   0.8314050
5    Alaska       Seattle     305    1841   0.14212488   0.8578751
6   AM West   Los Angeles     117     694   0.14426634   0.8557337
7   AM West       Phoenix     415    4840   0.07897241   0.9210276
8   AM West     San Diego      65     383   0.14508929   0.8549107
9   AM West San Francisco     129     320   0.28730512   0.7126949
10  AM West       Seattle      61     201   0.23282443   0.7671756
```

Note–Ontimeratio by definiton is the complement of Delayedratio–while it might not be necessary to calculate, I figure I thought I would include it for fun.

Statistical Comparison

Let's first compare statistical parameters for both Airlines

```
> flights %>% group_by(Airlines) %>% summarise(TotalDelay = sum(delayed))
# A tibble: 2 x 2
  Airlines TotalDelay
  <fct>         <int>
1 Alaska          501
2 AM West         787
```

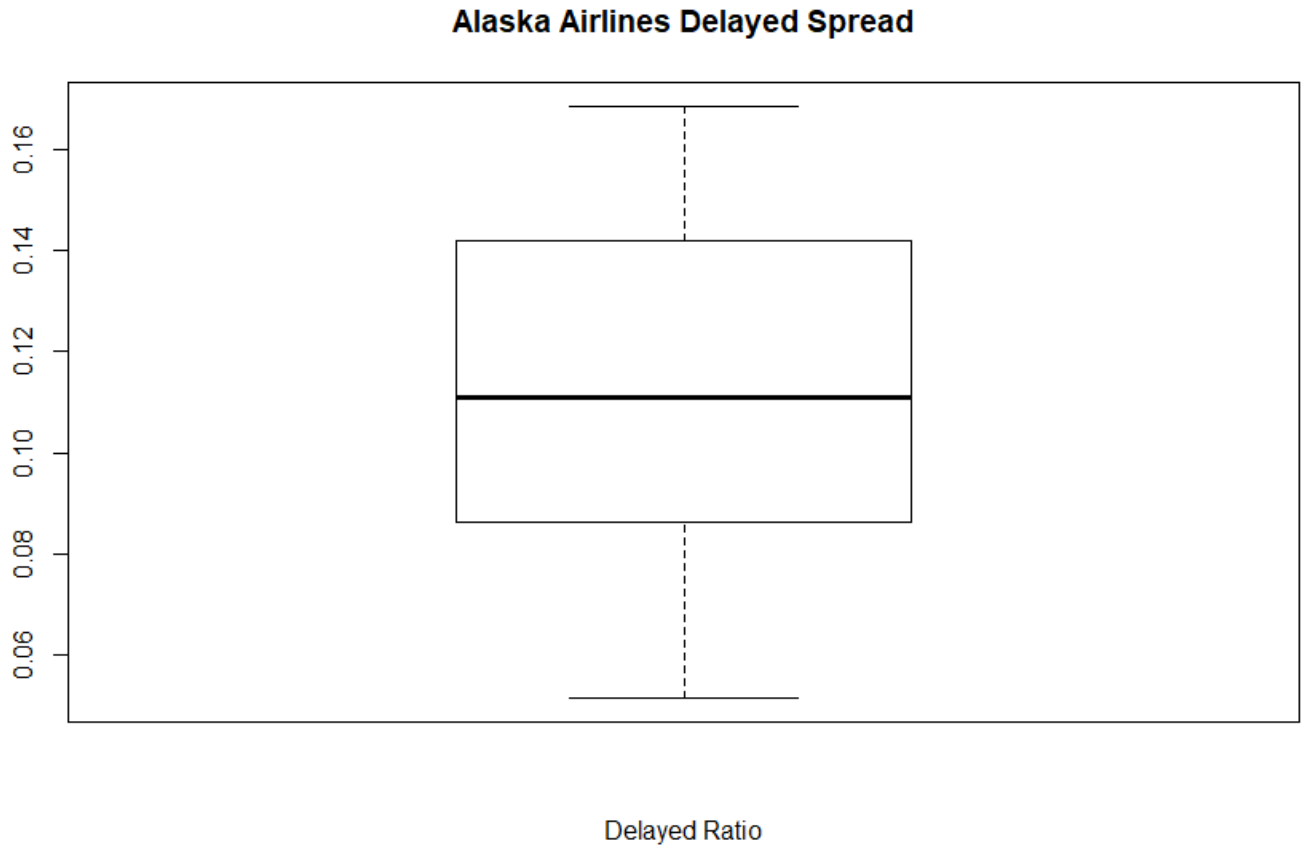
A head to head comparison shows that AM West has more delayed/ontime flights but that's due to just a higher total number of flights. Hence why I made a DelayedRatio (or %) metric for comparison,

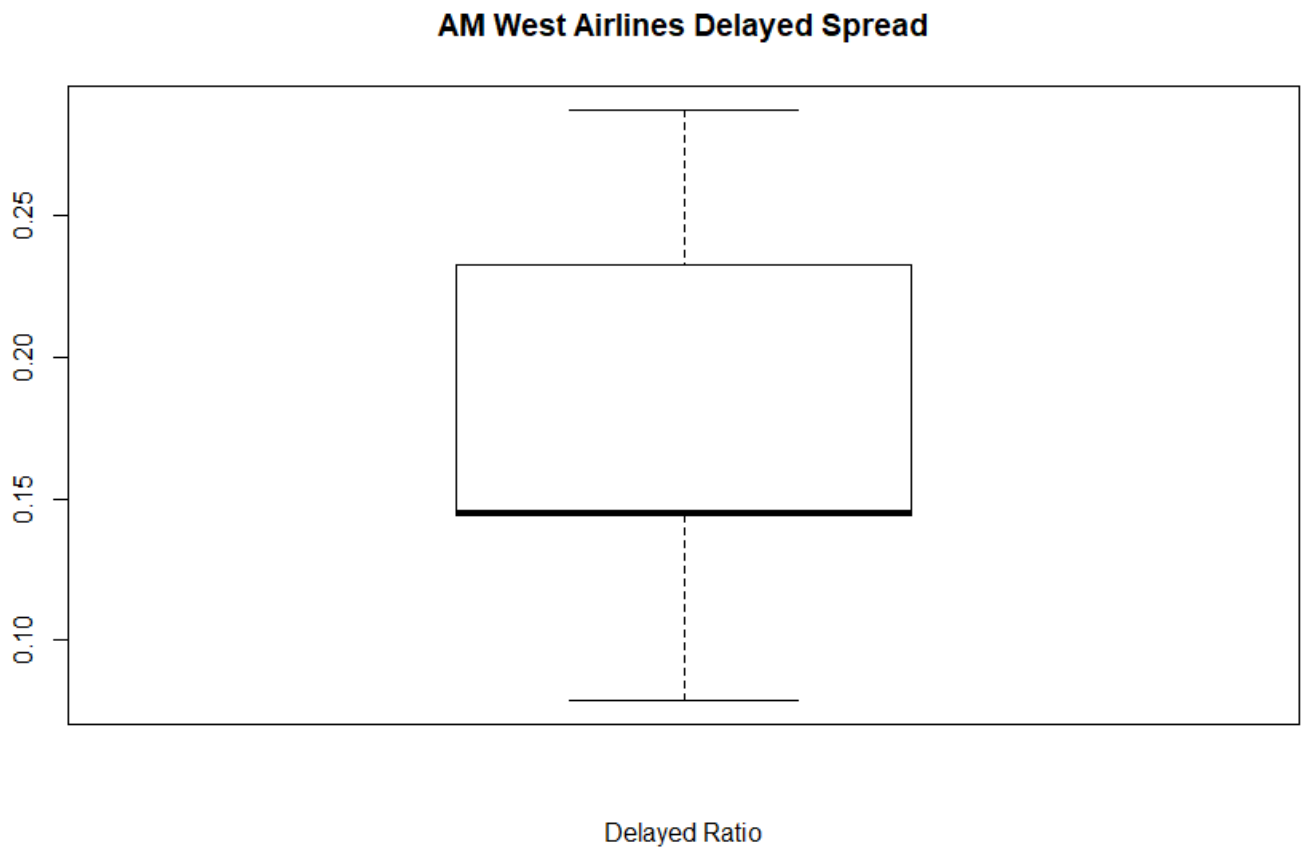
```
> summary((head(flights,5)))
    Airlines     City              delayed         on time        DelayedRatio      OnTimeRatio
        :0   Length:5           Min.   : 12.0   Min.   : 212.0   Min.   :0.05150   Min.   :0.831
4
 Alaska :5   Class :character   1st Qu.: 20.0   1st Qu.: 221.0   1st Qu.:0.08621   1st Qu.:0.857
9
 AM West:0   Mode  :character   Median : 62.0   Median : 497.0   Median :0.11091   Median :0.889
1
                                Mean   :100.2   Mean   : 654.8   Mean   :0.11187   Mean   :0.888
1
                                3rd Qu.:102.0   3rd Qu.: 503.0   3rd Qu.:0.14212   3rd Qu.:0.913
8
                                Max.   :305.0   Max.   :1841.0   Max.   :0.16860   Max.   :0.948
5
> summary(tail(flights,5))
    Airlines     City              delayed         on time      DelayedRatio      OnTimeRatio
        :0   Length:5           Min.   : 61.0   Min.   : 201   Min.   :0.07897   Min.   :0.7127
 Alaska :0   Class :character   1st Qu.: 65.0   1st Qu.: 320   1st Qu.:0.14427   1st Qu.:0.7672
 AM West:5   Mode  :character   Median :117.0   Median : 383   Median :0.14509   Median :0.8549
                                Mean   :157.4   Mean   :1288   Mean   :0.17769   Mean   :0.8223
                                3rd Qu.:129.0   3rd Qu.: 694   3rd Qu.:0.23282   3rd Qu.:0.8557
                                Max.   :415.0   Max.   :4840   Max.   :0.28731   Max.   :0.9210
```

Here we see that the entire DelayedRatio spread for Alaska air is lower than AM West–notice how the maximum delay on Alaska air is still less than the median of AM West; Alaska airlines has less delays overall than AM West–and by complemnet, a higher On Time percentage.

If you were randomly choosing a flight on an airline–you would want to fly Alaska due to less delays.

Just to reiterate–here are the boxplots





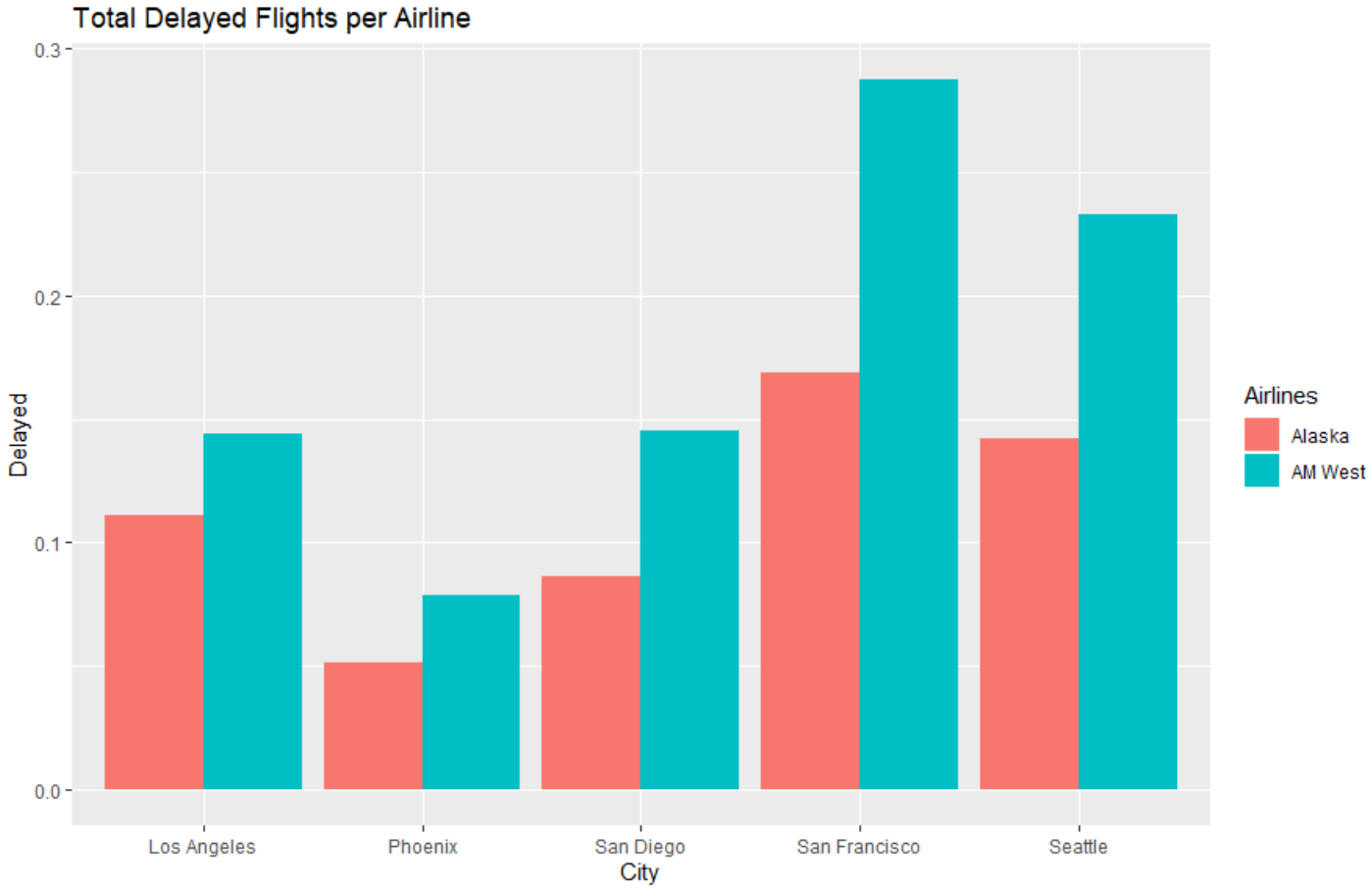
Notice the distinct lower median for Alaskan airline vs AM West

Comparison via Cities

So we know overall that Alaska has less delays than AM West but let's look at it on a city by city basis.

The easiest way to do this is just via good ole bar plot comparison.

```
> ggplot(flights, aes(x = City, y = DelayedRatio)) + geom_bar(aes(fill= Airlines), stat = "Ident
ity", position=position_dodge()) + ylab("Delayed %") + ggtitle("Delayed % Flights per Airline")
```



Airline Delay

So here we see the delays by city by Airlines–once again we see the Alaska delay %s much lower than that AM West–this was some what expected given the previous analysis.

Conclusions

From this plot we can make the following conclusions/comments: 1) Overall Alaska airlines have less delays across all cities when compared to AM West
2) San Francisco has the most delays (across both airlines) with Phoenix having the least
3) The largest magnitude difference in delays is via San Francisco–perhaps Alaska has some sort of advantage there, better gates?
4) Interesting–Los Angeles (the 2nd largest city in the US) had less delays per airline compared to smaller cities like San Francisco and Seattle.

This was a very useful excercise that showed the power and of tidying up dataframes for analysis.