hw5

Yan Lin

October 1, 2017

```
require(stringr)
```

```
## Loading required package: stringr
```

```
require(tidyr)
```

```
## Loading required package: tidyr
```

```
## Warning: package 'tidyr' was built under R version 3.4.2
```

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## 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
```

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
## read the csv table

a <- read.csv("C:/Users/Yan/Documents/flights.csv",header=TRUE, sep=",")
a
```

```
##         X     X.1 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
```

```
## set the blank cell as NA
a[a==""] <- NA

## clean the table
a[2,1] <- a[1,1]
a[5,1] <- a[4,1]
a <- na.omit(a)
head(a)
```

```
##         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
## 4 AM WEST on time         694    4840       383           320      201
## 5 AM WEST delayed         117     415        65           129       61
```

```
colnames(a)[1] <- "airline"
colnames(a)[2] <- "arrival"
head(a)
```

```
##   airline arrival 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
## 4 AM WEST on time         694    4840       383           320     201
## 5 AM WEST delayed         117     415        65           129      61
```

```
## tidy up data, 5 cities in a column and spread the arrival column into two
a <- a %>%
  gather(city, freq, 3:7) %>%
  spread(arrival, freq)
head(a)
```

```
##   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
```

```
## remove "." in the city name
a$city <- str_replace_all(a$city,"\\."," ")

## rename the colname "on time"
colnames(a)[4] <- "on_time"

## delay ratio per city per airline
a <- a %>%
    mutate(ratio = delayed / (delayed + on_time)) %>%
    arrange(desc(ratio))

## convert it to percentage
b <- a
b$ratio <- paste(round(100*b$ratio,2),"%",sep="")
head(b)
```

```
##   airline          city delayed on_time  ratio
## 1 AM WEST San Francisco     129     320 28.73%
## 2 AM WEST       Seattle      61     201 23.28%
## 3  ALASKA San Francisco     102     503 16.86%
## 4 AM WEST     San Diego      65     383 14.51%
## 5 AM WEST   Los Angeles     117     694 14.43%
## 6  ALASKA       Seattle     305    1841 14.21%
```

```
##delay ratio per city
a_city <- a %>%
    group_by(city) %>%
    summarise(average_delay = mean(ratio)) %>%
    arrange(desc(average_delay))
```

```
a_city$average_delay <- paste(round(100*a_city$average_delay,2),"%",sep="")
head(a_city)
```

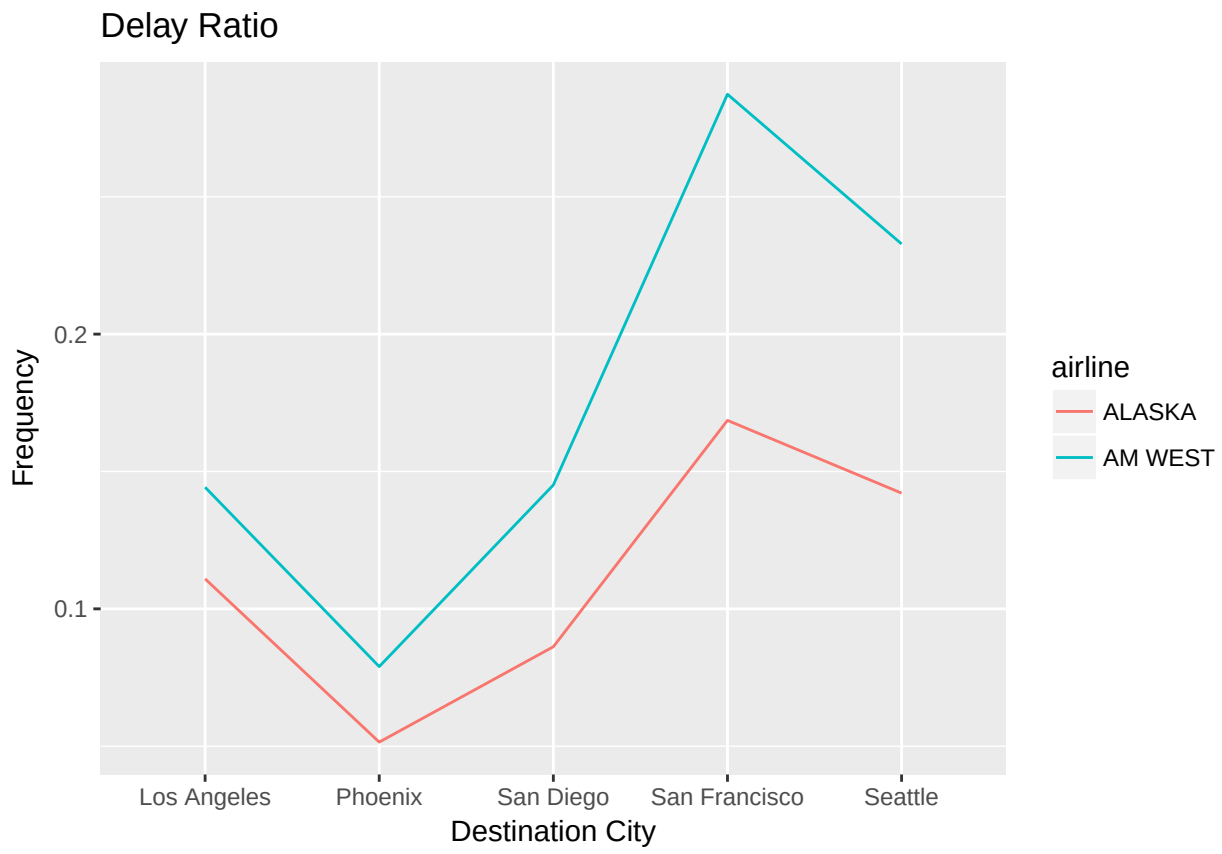
```
## # A tibble: 5 x 2
##            city average_delay
##           <chr>         <chr>
## 1 San Francisco         22.8%
## 2       Seattle        18.75%
## 3   Los Angeles        12.76%
## 4     San Diego        11.56%
## 5       Phoenix         6.52%
```

```
##delay ratio per airline
a_airline <- a %>%
    group_by(airline) %>%
    summarise(average_delay = mean(ratio)) %>%
    arrange(desc(average_delay))

a_airline$average_delay <- paste(round(100*a_airline$average_delay,2),"%",sep="")
head(a_airline)
```

```
## # A tibble: 2 x 2
##   airline average_delay
##    <fctr>         <chr>
## 1 AM WEST        17.77%
## 2  ALASKA        11.19%
```

```
##graph
g <- ggplot(a, aes( x= city, y = ratio ))
g <- g + geom_line(aes(color=airline,group = airline))
g <- g + labs(title = "Delay Ratio", x = "Destination City", y = "Frequency")
g
```




```
##conclusion
##The airline AM WEST has higher delay ratio than ALASKA
##throuhout the five cities, espcially in San Francisco,
##the gap is near to 12%. Both airlines perform best in Phoenix.
##Despite the ratio difference among these two airlines,
##the trend is similar. One might wnat to take a close look
##of the airports to determine what caused the delay.
```