HW5_607 tidyr & dplyr

Chunhui Zhu

October 1, 2017

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.

```
library("tidyr")
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
```

(2) Read the information from your .CSV file into R, and use tidyr and dplyr as needed to tidy and transform your data.

```
#read data from .csv file

df<-data.frame(read.csv("hflight.csv",stringsAsFactors = FALSE))
df
```

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

```
#fill in missed data in df$X

df[2,1]<-"ALASKA"
df[5,1]<-"AMWEST"
df
```

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

```
#filter function remove the row which x.1 is empty, read in from second column

df<-filter(df,df$X!="")
df
```

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

```
#gather number by airline and status

df2<- gather(df, key = "X.1", value = "count", Los.Angeles:Seattle)
colnames(df2)<-c("airline","status","city","number")
df2
```

```
##    airline  status          city number
## 1   ALASKA on time   Los.Angeles    497
## 2   ALASKA delayed   Los.Angeles     62
## 3   AMWEST on time   Los.Angeles    694
## 4   AMWEST delayed   Los.Angeles    117
## 5   ALASKA on time       Phoenix    221
## 6   ALASKA delayed       Phoenix     12
## 7   AMWEST on time       Phoenix   4840
## 8   AMWEST delayed       Phoenix    415
## 9   ALASKA on time     San.Diego    212
## 10  ALASKA delayed     San.Diego     20
## 11  AMWEST on time     San.Diego    383
## 12  AMWEST delayed     San.Diego     65
## 13  ALASKA on time San.Francisco    503
## 14  ALASKA delayed San.Francisco    102
## 15  AMWEST on time San.Francisco    320
## 16  AMWEST delayed San.Francisco    129
## 17  ALASKA on time       Seattle   1841
## 18  ALASKA delayed       Seattle    305
## 19  AMWEST on time       Seattle    201
## 20  AMWEST delayed       Seattle     61
```

```
#reorder table base on status

df2[order(df2$airline,df2$status),]
```

```
##    airline  status          city number
## 2   ALASKA delayed   Los.Angeles     62
## 6   ALASKA delayed       Phoenix     12
## 10  ALASKA delayed     San.Diego     20
## 14  ALASKA delayed San.Francisco    102
## 18  ALASKA delayed       Seattle    305
## 1   ALASKA on time   Los.Angeles    497
## 5   ALASKA on time       Phoenix    221
## 9   ALASKA on time     San.Diego    212
## 13  ALASKA on time San.Francisco    503
```

```
## 17  ALASKA on time       Seattle   1841
## 4   AMWEST delayed   Los.Angeles    117
## 8   AMWEST delayed       Phoenix    415
## 12  AMWEST delayed     San.Diego     65
## 16  AMWEST delayed San.Francisco    129
## 20  AMWEST delayed       Seattle     61
## 3   AMWEST on time   Los.Angeles    694
## 7   AMWEST on time       Phoenix   4840
## 11  AMWEST on time     San.Diego    383
## 15  AMWEST on time San.Francisco    320
## 19  AMWEST on time       Seattle    201
```

```
#split table base on the airline and status
#totally divid into four tables

st<-split(df2, with(df2, interaction(airline,status)), drop = TRUE)
st$ALASKA.delayed
```

```
##    airline  status          city number
## 2   ALASKA delayed   Los.Angeles     62
## 6   ALASKA delayed       Phoenix     12
## 10  ALASKA delayed     San.Diego     20
## 14  ALASKA delayed San.Francisco    102
## 18  ALASKA delayed       Seattle    305
```

```
st$`ALASKA.on time`
```

```
##    airline  status          city number
## 1   ALASKA on time   Los.Angeles    497
## 5   ALASKA on time       Phoenix    221
## 9   ALASKA on time     San.Diego    212
## 13  ALASKA on time San.Francisco    503
## 17  ALASKA on time       Seattle   1841
```

```
st$AMWEST.delayed
```

```
##    airline  status          city number
## 4   AMWEST delayed   Los.Angeles    117
## 8   AMWEST delayed       Phoenix    415
## 12  AMWEST delayed     San.Diego     65
## 16  AMWEST delayed San.Francisco    129
## 20  AMWEST delayed       Seattle     61
```

```
st$`AMWEST.on time`
```

```
##    airline  status          city number
## 3   AMWEST on time   Los.Angeles    694
## 7   AMWEST on time       Phoenix   4840
## 11  AMWEST on time     San.Diego    383
## 15  AMWEST on time San.Francisco    320
## 19  AMWEST on time       Seattle    201
```

3. compare the arrival delays for the two airlines.

```
#calculat delay rate for airlines
#built data frame t to store the data set
```

```
Alaska_delay_r<-st$ALASKA.delayed$number/(st$ALASKA.delayed$number+st$`ALASKA.on time`$number)
Amwest_delay_r<-st$AMWEST.delayed$number/(st$AMWEST.delayed$number+st$`AMWEST.on time`$number)

cities<-c("Los.Angeles","Phoenix","San.Diego","San.Francisco","Seattle")
t<-data.frame(cities, round(Alaska_delay_r,2), round(Amwest_delay_r,2))
t
```

```
##          cities round.Alaska_delay_r..2. round.Amwest_delay_r..2.
## 1   Los.Angeles                     0.11                     0.14
## 2       Phoenix                     0.05                     0.08
## 3     San.Diego                     0.09                     0.15
## 4 San.Francisco                     0.17                     0.29
## 5       Seattle                     0.14                     0.23
```

```
#The result shows Amwest has higher delay rate than Alaska in 5 cities
#Difference for taking Alaska to LA,Phonenix,San.Diego,San.Francisco,Seattle,
#you will probably have lesser risk to delay.
```

Another thought for analysis

```
#We lack of infomation about collected data like by seasons,daytime,
#passenger group,from which cities, and etc.
#Even though the result from the above result shows Amwest airline
#having more delay, it doesn't convince people taking Amwest airline
#from city like NY will get more chance to delay than taking Alaska airline.
```

```
#creat a airlines.csv file for t data set at "C:/Users/Ivy/Desktop/607/W5"

setwd("C:/Users/Ivy/Desktop/607/W5")
write.csv(t,"t.csv")
```

(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.

Please check out two URLs from the blackboard.