Tidying and Transforming Data

Sharon Morris

September 25, 2016

(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.

Import csv file

```
week6 <- read.csv('week5.csv',header=TRUE,stringsAsFactors = FALSE)
week6
```

```
##         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
```

Remove all rows and missing data

```
library(tidyr)

na <- na.omit(week6)
head(na)
```

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

Add airline name to rows

Without the added row names the spread will not run because values in rows without names are treated as duplicates

```
na[2,1] <-  na[1,1]
na[4,1] <-  na[3,1]

head(na)
```

```
##         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
```

Insert column names

```
names(na)[1:2] <- c('Airline', 'Status')
na
```

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

Transform data – make city into 1 column

```
na <- gather(na, "City", Count,3:7)
na
```

```
##    Airline  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
```

Remove . from state names

```
na$City<-gsub("\\."," ",na$City)

na
```

```
##    Airline  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
```

Separate values in Status column

```
na<-spread(na, Status, Count)

head(na)
```

```
##   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
```

Rename column on time

```
colnames(na)[4] <- "Ontime"

head(na)
```

```
##   Airline          City delayed Ontime
## 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
```

Analysis of airline arrival and departures

```
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
```

```
Compare <- na %>% group_by(City, Airline) %>% summarise(ComOntime=sum(Ontime),
                                      ComDelay=sum(delayed),
                                    Compariosn=ComOntime/(ComOntime+ComDelay)*100)
Compare
```

```
## Source: local data frame [10 x 5]
## Groups: City [?]
##
##             City Airline ComOntime ComDelay Compariosn
##            <chr>   <chr>     <int>    <int>      <dbl>
## 1    Los Angeles  ALASKA       497       62   88.90877
## 2    Los Angeles AM WEST       694      117   85.57337
## 3        Phoenix  ALASKA       221       12   94.84979
```

```
## 4        Phoenix AM WEST      4840      415   92.10276
## 5      San Diego  ALASKA       212       20   91.37931
## 6      San Diego AM WEST       383       65   85.49107
## 7  San Francisco  ALASKA       503      102   83.14050
## 8  San Francisco AM WEST       320      129   71.26949
## 9        Seattle  ALASKA      1841      305   85.78751
## 10       Seattle AM WEST       201       61   76.71756
```