DATA607 Assignment 4: Tidying and Transformation Data

Albert Gilharry

27 February 2018

Contents

Intro

For this assignment we were given a data set containing on time and delay information of 2 airlines. We were tasked with tidying, transforming, and analyzing the data. My solution is documented below

Load Libraries

```
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
```

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

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
airline,performance,Los Angeles,Phoenix,San Diego,San Francisco,Seattle
ALASKA,on time,497,221,212,503,1841
ALASKA,delayed,62,12,20,102,305
AM WEST,on time,694,4840,383,320,201
AM WEST,delayed,117,415,65,129,61
```

Figure 1:

My CSV

Below is an image of the CSV file I created from the table in the assignment document.

Load CSV

```
raw_data <- read.csv("data/data.csv", sep = ",", header = TRUE, stringsAsFactors=FALSE)
print(raw_data)
```

```
##   airline performance 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 Data

Use the gather function to have the city as rows, set have their respective flights as a counts column.

```
raw_data = gather(raw_data, "city", "count", 3:7)
print(raw_data)
```

```
##    airline performance          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
```

Use the `spread` function to create `ontime` and `delayed` columns. This is makes it easier to visualize the data.

```
raw_data = spread(raw_data, "performance", "count")
print(raw_data)
```

```
##    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
## 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
```

Analysis: Which Airline Performed Better?

To find out, lets answer the following questions?

- What are their respective market shares?

```
total_flights =  sum(raw_data$`on time`) + sum(raw_data$`delayed`)
mutate(raw_data, flights = delayed + `on time`) %>%
  group_by(airline) %>%
    summarise(airline_total=(sum(flights)/total_flights))
```

```
## # A tibble: 2 x 2
##   airline airline_total
##   <chr>           <dbl>
## 1 ALASKA          0.343
## 2 AM WEST         0.657
```

AM WEST has an advantageous 65.7% market share when compared to ALASKA at only 34.3%.

- Do the market shares differ across cities?

This may be better shown by a plot.

```
      ggplot(data = mutate( raw_data, flights = delayed + `on time`) %>%
               group_by(city,airline) %>%
               summarise(airline_total=(sum(flights))),
             aes(x = city, y = airline_total,fill=airline)) +
                geom_bar(stat = "identity") +
                coord_flip() +
                labs(title = "Airlines Market Share by City")
```

The plot shows that AM WEST has a larger market share in 3 of the 5 cities and it dominates in Phoenix. ALASKA dominates in Seattle but at a significantly lower magnitude than AM WEST's domination in Phoenix.

- What is each airline overall on time ratio?

```
  select(raw_data, -city) %>%
  group_by(airline) %>%
    summarise(`on time rate`=( sum(`on time`)/( sum(`on time`) + sum(`delayed`) ) ) )
```

```
## # A tibble: 2 x 2
##   airline `on time rate`
##   <chr>            <dbl>
## 1 ALASKA           0.867
## 2 AM WEST          0.891
```

AM WEST has a higher on time rate of 89.1% when compared to ALASKA at 86.7%

- Would you consider choosing an airline based on your destination?

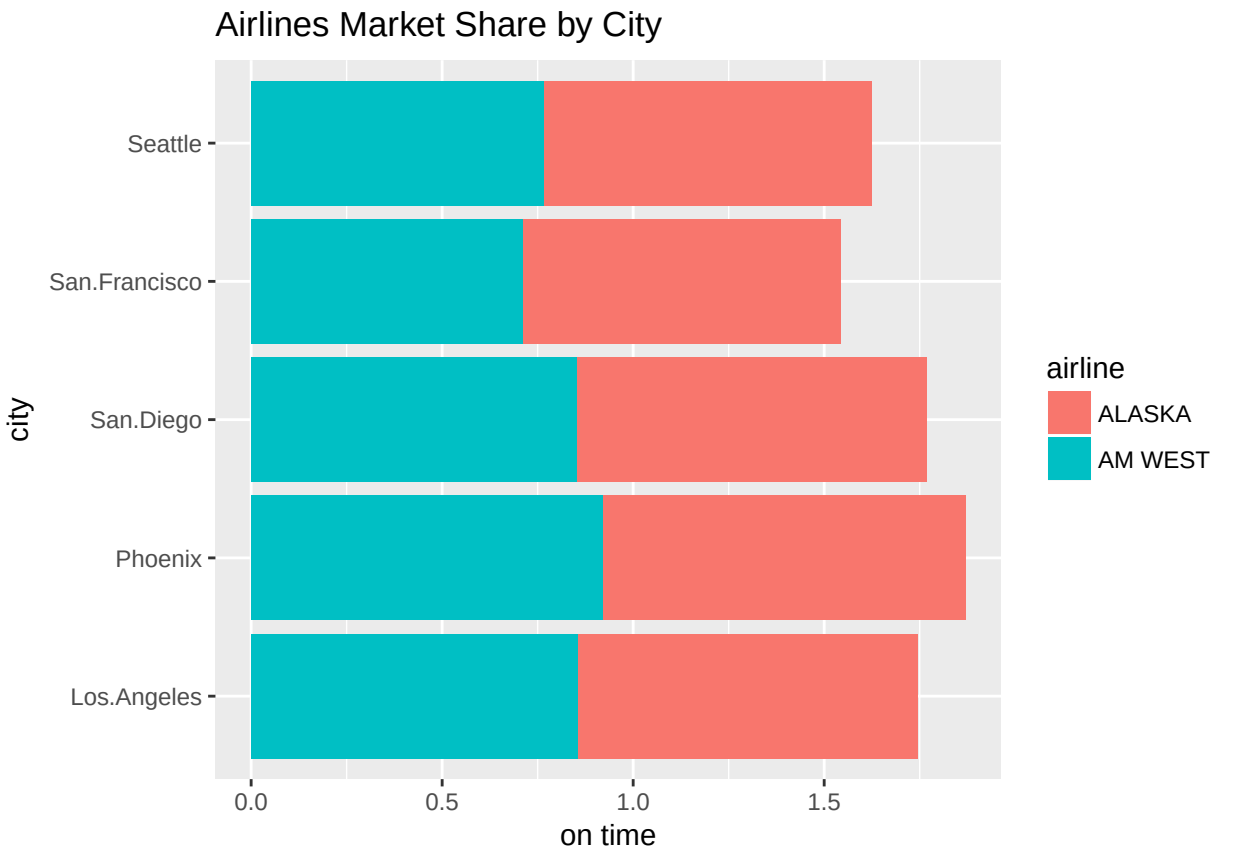
To answer this, lets compare the on time performance of airlines by city.

Firstly, lets do a visual.

```
 data =  select(raw_data,everything()) %>%
  group_by(airline,city) %>%
    summarise(`on time`=( sum(`on time`)/( sum(`on time`) + sum(`delayed`) ) ) ) %>%
    arrange(city)

ggplot(data = data, aes(x = city, y = `on time`,fill=airline)) +
```

```
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Airlines Market Share by City")
```



Strangely it seems the ALASKA performs better, we need to look at the data to be sure.

```
print(spread(data,city,`on time`))
```

```
## # A tibble: 2 x 6
## # Groups:   airline [2]
##   airline Los.Angeles Phoenix San.Diego San.Francisco Seattle
##   <chr>         <dbl>   <dbl>     <dbl>         <dbl>   <dbl>
## 1 ALASKA        0.889   0.948     0.914         0.831   0.858
## 2 AM WEST       0.856   0.921     0.855         0.713   0.767
```

This is interesting! Alaska has a higher on time rate in every city, despite AM WEST having a higher overall on time rate! ALASKA's on time rate advantage in San Diego, San Francisco, and Seattle ranges from 5.9% to 11.8%, a significant advantage. AM WEST's severely dominant market share in Phoenix skewed their overall on time rate and it misled us to believe that they are better.

The data suggests that ALASKA is the better airline (based on on-time rates), despite having a lower market share. This highlights the importance of exploratory data analysis, it reveals insights that are not obvious by simply skimming through the data!