Tidying and Transforming Data

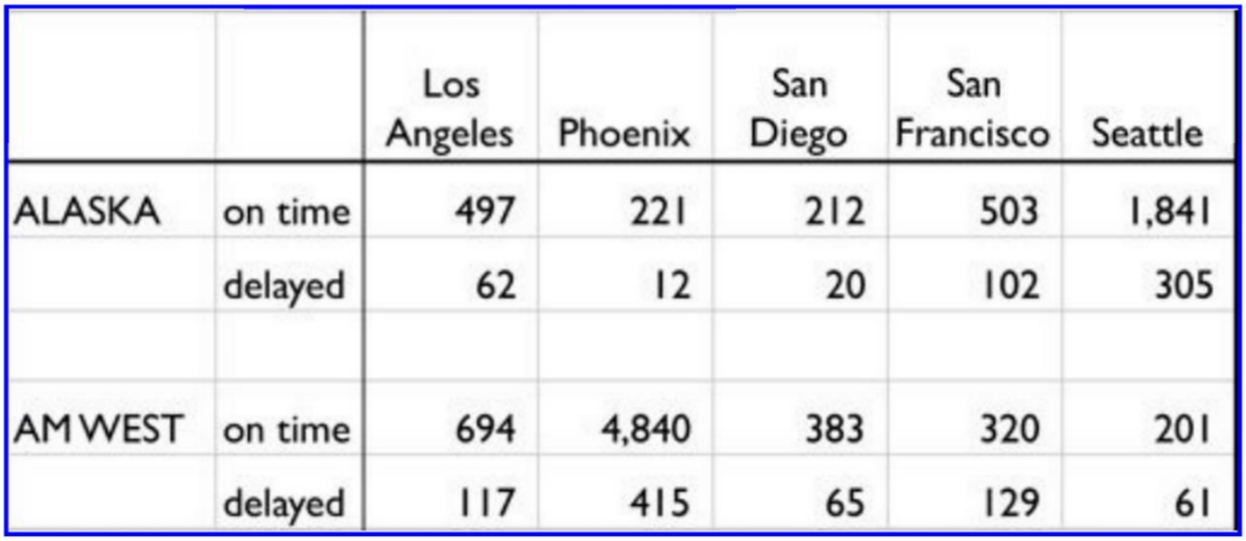
Z. Palmore

9/25/2020

Assignment 4

```r
library(tidyverse)
library(readr)
library(reshape2)
```

Directions



		Los Angeles	Phoenix	San Diego	San Francisco	Seattle
ALASKA	on time	497	221	212	503	1,841
	delayed	62	12	20	102	305
AM WEST	on time	694	4,840	383	320	201
	delayed	117	415	65	129	61

Source: *Numbersense*, Kaiser Fung, McGraw Hill, 2013

Figure 1: AirlineArrivals

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.

(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.

(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.

Step 1: Creating the CSV

A .csv file was created by manually entering the records from the chart into a spreadsheet called “Sample_Flights.” The file includes all data from the chart as it was formatted. It can be viewed in a separate window by running this chunk.

```
setwd("C:/Users/Owner/Documents/GitHub/msdsdata607/Assignments/Assignment 4")
sampleflights <- read_csv("Sample_Flights.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1], 'X2' [2]
```

```
# sampleflights <- read.csv("GitHub/msdsdata607/Assignments/Assignment 4/Sample_Flights.csv")
View(sampleflights)
```

Step 2: Importing and Tidying

With data written into the .csv file and imported exactly as was in the chart, there are many NA values spread throughout the data. These should be removed and the data cleaned to make analysis easier.

For background, the data contains on-time and delayed arrivals for two airlines Alaska and AM West. There are 159 observations, 5 rows (currently) and 7 variables which include the airline name, its status (of on-time or delayed), and the names of locations from 5 selected airports where the airline arrived. The airport locations are Los Angeles (LAX), Phoenix (PHX), San Diego (SAN), San Francisco (SFO), and Seattle (SEA).

Prior to analysis it would be best to complete these tasks:

- Add Column Names for;
  - Airline
  - Status
- Add the Airline names for
  - delayed status of
    - Alaska
    - AM West
- Remove the 3rd Row
  - Containing all NA values

This was accomplished using the following steps.

```
# LAX, PHX, SAN, SFO, & SEA
# Creating a data from of the data
sampleflights <- as.data.frame(sampleflights)
# Specify the column names needed
colnames(sampleflights) <- c("Airline",
                             "Status",
                             "LAX",
                             "PHX",
                             "SAN",
                             "SFO",
                             "SEA")
# Remove row 3 since it is all NA's
sampleflights <- sampleflights[-3,]
# Add the character string "Alaska" where NA is listed
sampleflights[2,1] <- ("Alaska")
# Add the character string "AM West" where NA is listed
sampleflights[4,1] <- ("AM West")
# Review the data frame
sampleflights
```

```
##   Airline  Status LAX  PHX SAN SFO  SEA
## 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
```

Now there are four rows in the data since the row that was full of NA's was removed. The proper names were given to the row with delayed data for each airline and the column names for airline and its status. All other information in the data frame remains the same.

Step 3: Arrival Delay Analysis

There are several ways to compare the two airlines' arrival delays depending on the type of units recorded with this data. From my perspective, there are two options in this set. Data could be, for example, recorded in a quantity of minutes or quantity of flights. Regardless, the first step is to combine all arrival delays for each location. Another is to find the mean of each row across those same locations. One could also use these to find the probability of delay for each airline. All will be performed for comparison.

```
# This code sums the rows using the all data from columns 3 through 7
# Then it creates a new column in the data frame
# which contains the sums of each row called "Total"

sampleflights$Total <- rowSums(sampleflights[,3:7])
sampleflights$Means <- rowMeans(sampleflights[,3:7])

# Review the data frame with the airline, status, and new columns
sampleflights[,c(1,2,8,9)]
```

```
##   Airline  Status Total  Means
## 1  Alaska on time  3274  654.8
## 2  Alaska delayed   501  100.2
## 4 AM West on time  6438 1287.6
## 5 AM West delayed   787  157.4
```

Based on this analysis, the airline AM West has the largest quantity of arrival delays in both its total delayed and its average delayed. However, it also has more flights than Alaska overall which can be misleading when considering the full picture. For now, we are only reviewing raw numbers.

Given that there were only two airlines, AM West's higher amount of arrival delays makes Alaska's airline the least delayed by the quantity of data alone. Since this airline (AM West) has the highest amount of arrival delays, the only remaining airline must have the shortest delayed time (Alaska). If minutes were used to measure and record delays, then it is also clear both airlines had average delays greater than 100 minutes.

Other statistics show the minimum delay occurred with Alaska when arriving in the location of Phoenix at 12. The maximum arrival delay also occurred in Phoenix but with the Airline AM West. This is shown here using the Minima and Maxima columns that were created for this purpose.

```r
# Applying the minimum function over the rows (by using the margin of 1) in the data frame
# then storing results in a new column called "Minima"
sampleflights$Minima <- apply(sampleflights[3:7], 1, FUN = min)

# Selecting the variables to display the minimum flight, its status and the airline
sampleflights[2,c(1,2,4)]
```

```
##   Airline  Status PHX
## 2  Alaska delayed  12
```

This displayed the minimum flight delay for the study at 12.

```r
# Repeat the same function for the maximum of flights and calling the column "Maxima"
sampleflights$Maxima <- apply(sampleflights[,3:7], 1, FUN = max)

# Selecting the variables to display the maximum flight, its status and the airline
sampleflights[3,c(1,2,4)]
```

```
##   Airline  Status  PHX
## 4 AM West on time 4840
```

Then this displayed the maximum delay at 4840. Now instead of only comparing raw data, let's consider the numbers in context. If Alaska had far fewer flights than AM West, then it is not realistic to immediately say one is less delayed than the other without a proper comparison of the two arrival delays overall.

The total on-time flights with Alaska and AM West were 3274 and 6438 respectively. The total delayed for each were 501 and 787 respectively. If the records directly indicated the quantity of flights that were on-time or delayed and we assume that a flight cannot be both on-time and delayed, then we can create a ratio of delayed to on-time flights as shown here;

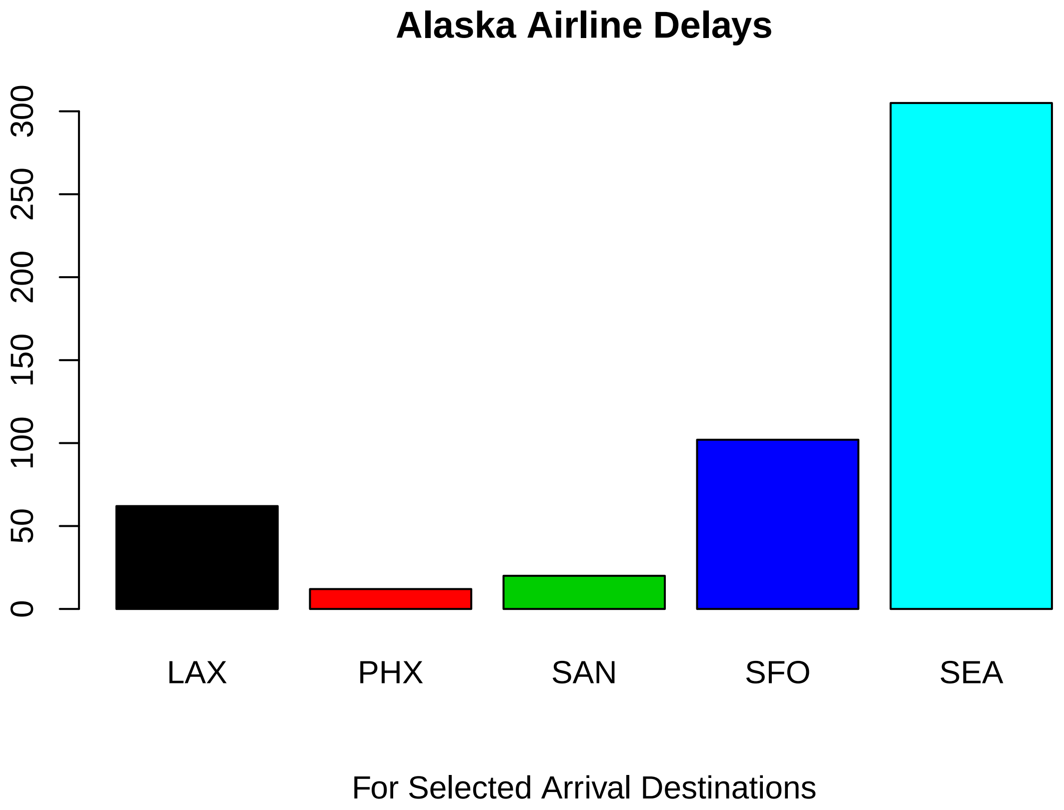
```r
# Sum of flights for Alaska
ak_totalflights <- sampleflights[2,8] + sampleflights[1,8]
# Sum of flights for AM West
am_totalflights <- sampleflights[4,8] + sampleflights[3,8]
# Probability of delay for airline Alaska
AK_ratio <- sampleflights[2,8] / ak_totalflights
# Probability of delay for airline AM West
AM_ratio <- sampleflights[4,8] / am_totalflights
# Rounding AK_ratio to the correct number of significant figures, 3
PAK <- signif(AK_ratio, digits = 3)
# Rounding AK_ratio to the correct number of significant figures, 3
PAM <- signif(AM_ratio, digits = 3)
```

This calculation results in the probability of being delayed for each airline based on the selected flights for this dataset. We can see that the probability of delay on airline Alaska is 0.133 or about 13.3%. For AM West, the probability of delay is 0.109 or about 10.9%. Using this method of analysis, AW West has a slightly lower chance of being delayed based on the ratio of its total delays 787 to its total number of flights at 7225. While Alaska's chance of delay is slightly higher, it also had fewer flights at 501 delayed to 3775 flights total. It may also be helpful to graph the variables and demonstrate the differences.

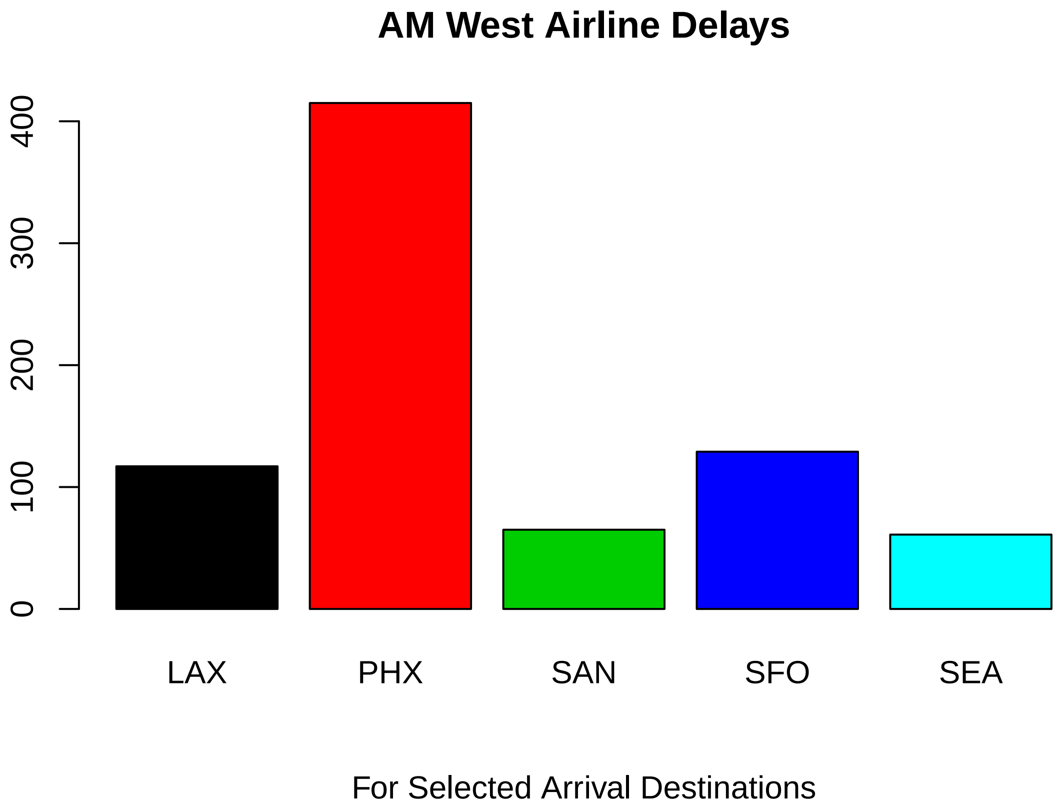
```r
# Create a new data frame with columns for plotting in ggplot
someflights <- melt(sampleflights)
```

```
## Using Airline, Status as id variables
```

```r
delays <- someflights %>%
  filter(Status == "delayed")
# Drop the statitics that are not needed by selecting what is
delays <- delays[1:10,]
# Creating a data frame with delays for Alaska
AK_delays <- delays %>%
  filter(Airline == "Alaska")
# Creating a data frame with delays for AW West
AM_delays <- delays %>%
  filter(Airline == "AM West")
# plotting delays for each location
barplot(AK_delays$value, horiz = FALSE, # unfortunately it displays poorly if TRUE
        names.arg = AK_delays$variable,
        main = "Alaska Airline Delays",
        sub = "For Selected Arrival Destinations",
        col = AK_delays$variable)
```

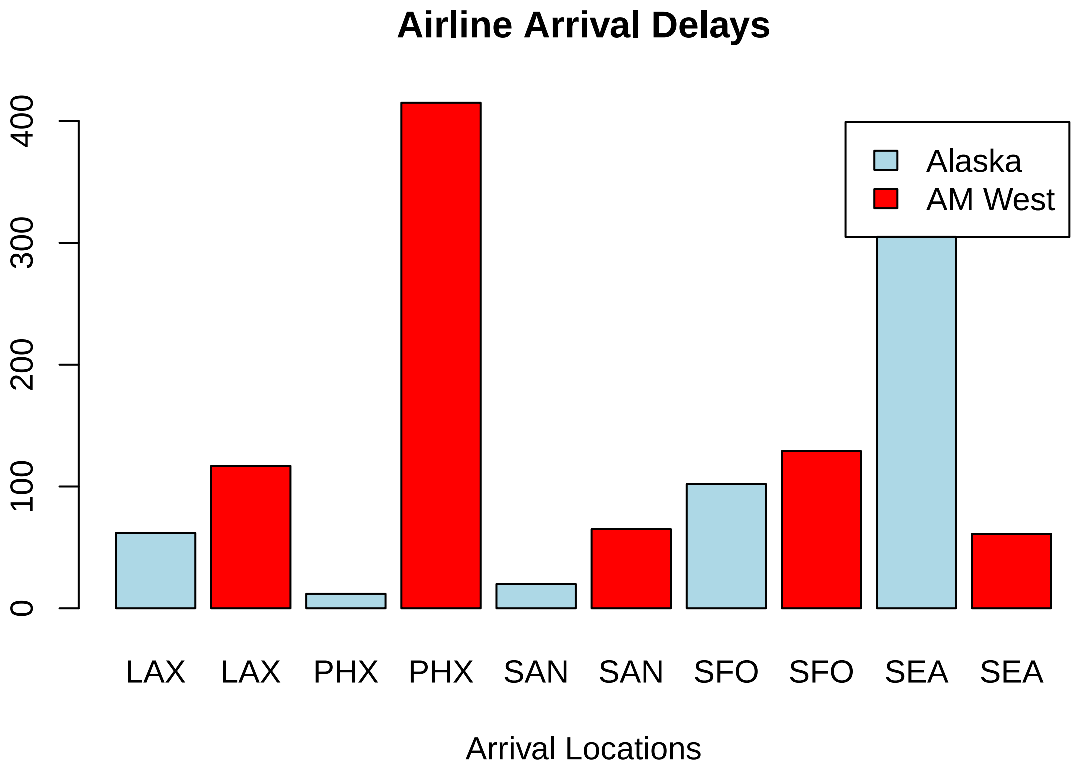


```r
barplot(AM_delays$value, horiz = FALSE,
        names.arg = AM_delays$variable,
        main = "AM West Airline Delays",
        sub = "For Selected Arrival Destinations",
        col = AM_delays$variable)
```



These bar charts show the difference in arrival delays for each airline by location. It is worth noting the difference in the x-axis where AM West has enough flights (or flight time) to increase the range displayed.

```
barplot(delays$value,
        col = c("light blue", "red"),
        main = "Airline Arrival Delays",
        names.arg = delays$variable,
        xlab = "Arrival Locations",
        legend = c("Alaska","AM West"),
        beside = TRUE)
```

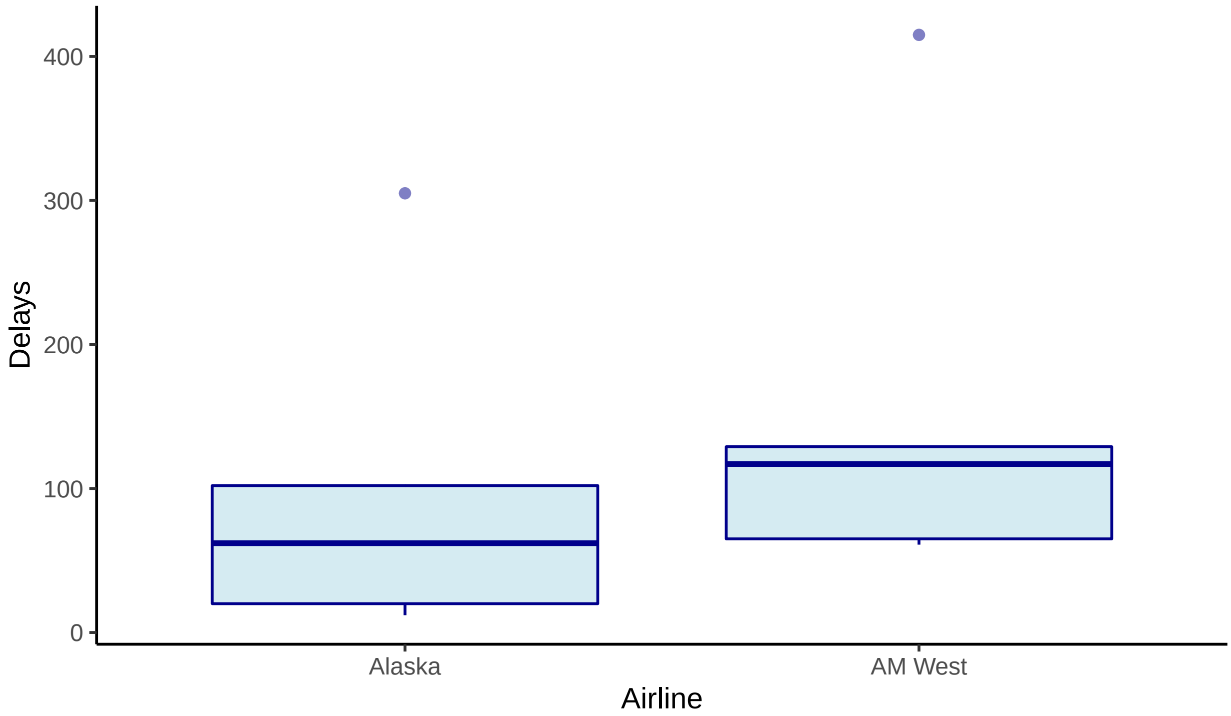


This bar chart demonstrates that difference in total number of flights a little better. Across the chart we can see AM West almost has consistently more flights than Alaska. It is also worth noting that both seem to be focusing on arrivals at one location. AM West has most of its flights arriving in second blue bar (which happens to be Phoenix and the largest bar), while Alaska concentrates its flights in the last red bar (of Seattle). The difference in distributions however, can be shown in the boxplots below.

```
ggplot(data = delays, aes(x = Airline, y = value) ) +
  geom_boxplot(color="Dark Blue", fill="Light Blue", alpha=0.5) +
  labs(title = "Distribution of Airline Delays", subtitle = "With Selected Airlines",
       y ="Delays", caption = "Source: Numbersense, Kaiser Fung, McGraw Hill, 2013") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5),
        plot.caption = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
```

Distribution of Airline Delays

With Selected Airlines



Source: Numbersense, Kaiser Fung, McGraw Hill, 2013

There are two points that are skewing the data. One at about 300 on the y-axis for airline Alaska and another around 400 for AM West. Their interquartile ranges are visibly different, with Alaska's being larger and centered (and spaced) evenly about its mean. However, AM West is not centered, it is heavily skewed towards higher delays and thus its mean is much higher in its "box." These distributions may help explain why we can run two different methods of analysis (one using raw values and another using calculated probability) and get two different results.

Submission

This document has been submitted to rpubs and is also contained within this GitHub repository.